Journal of Biomedical Informatics 101 (2020) 103323

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

SECNLP: A survey of embeddings in clinical natural language processing

Katikapalli Subramanyam Kalyan*, S. Sangeetha

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

ARTICLE INFO

Keywords:
Embeddings
Distributed representations
Medical
Natural language processing
Survey

ABSTRACT

Distributed vector representations or embeddings map variable length text to dense fixed length vectors as well as capture prior knowledge which can transferred to downstream tasks. Even though embeddings have become de facto standard for text representation in deep learning based NLP tasks in both general and clinical domains, there is no survey paper which presents a detailed review of embeddings in Clinical Natural Language Processing. In this survey paper, we discuss various medical corpora and their characteristics, medical codes and present a brief overview as well as comparison of popular embeddings models. We classify clinical embeddings and discuss each embedding type in detail. We discuss various evaluation methods followed by possible solutions to various challenges in clinical embeddings. Finally, we conclude with some of the future directions which will advance research in clinical embeddings.

1. Introduction

Distributed vector representation or embedding is one of the recent as well as prominent addition to modern natural language processing. Embedding has gained lot of attention and has become a part of NLP researcher's toolkit. Representations based on word frequency, tf-idf measure, N-grams etc. are high dimensional, sparse and ignore order as well as syntactic and semantic similarities of the words. In contrast, embedding maps variable length text to dense vector representations and overcome issues like a) curse of dimensionality and b) lack of syntactic and semantic information in representations. Moreover, embeddings are learned in an unsupervised manner which capture knowledge in large unlabeled corpus and the captured knowledge can be transferred to downstream tasks with small labeled data sets. Hence, embedding has become an unavoidable choice for text representation in recent times of deep learning era.

Research in learning distributed vector representations started with [1] and then several research studies [2–8] laid foundation for research in embeddings. Word2Vec proposed by [9] brought immense popularity to embeddings and then models like glove [10], fastText [11], ELMo [12] and BERT [13] were proposed. Even though embeddings have become de facto standard for text representation in deep learning based NLP tasks in both general and medical domains, there is no survey paper which presents a detailed review of embeddings in the form of classification of embeddings as well as the challenges to be solved. To the best of our best knowledge, we are the first to present a detailed review of embeddings in Clinical Natural Language Processing.

1.1. Literature Selection

We collected papers from various sources like PubMed, Google Scholar, ScienceDirect, ACL Web Anthology, and AAAI.We confined to the papers which are published in the period January 2014 to Nov 2018 because of the recent popularity of embeddings. We used keywords like "deep learning," "medical", "clinical," "embeddings", "natural language processing", "distributed representations" and "health" to retrieve the relevant papers and gathered 230 articles. After the removal of duplicate articles as well as the articles which are not related to clinical natural language processing, the number of articles reduced to 120. Finally, we included the most relevant 80 papers after a manual review of all the remaining articles. We summarize key contributions of our survey article as

- First attempt to provide a comprehensive review of embeddings in clinical natural language processing.
- We classify medical corpora into four types depending on the source, discuss each type in detail and finally provide comparison to highlight characteristics of each corpus type. (Section 2)
- We briefly discuss various medical codes as well as their significance. (Section 3)
- We discuss popular embedding models like word2vec, glove, fastText, doc2vec, ELMo and then provide comparison to highlight advantages and disadvantages of each. (Section 4)
- We classify embeddings into two types depending on whether they map text or concepts. Further, text embeddings are classified into five types depending on the granularity of text they map and

* Corresponding author.
E-mail addresses: kalyan.ks@yahoo.com (K.S. Kalyan), sangeetha@nitt.edu (S. Sangeetha).

https://doi.org/10.1016/j.jbi.2019.103323
Received 17 March 2019; Received in revised form 12 September 2019; Accepted 27 October 2019
Available online 08 November 2019

concept embeddings are classified into three types depending on the concept (code, CUI or patient) they map. (Section 5)
- We discuss various methods like intrinsic and extrinsic to evaluate embeddings and present summary of evaluation tasks in various clinical embeddings. (Section 6)
- We discuss various challenges like *small size of clinical corpus*, *multi sense embeddings*, *domain adaptation*, *sub-word information*, *OOV issue*, *temporal information* and suggest possible solutions from the surveyed research articles. (Section 7)
- Finally, we conclude with some of future directions of research in embeddings like *interpretability*, *knowledge distillation*, *bias* and *evaluation of embeddings*. (Section 8)

2. Medical corpora

In this section, we classify medical corpora into four types as shown in Fig. 1 and then discuss each type followed by a comparison (see Table 1).

Embeddings are inferred using any of the embeddings models over a large unlabeled corpus. Quality of embeddings inferred, depends on two properties of corpus like size and whether it is general or domain specific. A large corpus provides better vocabulary coverage while a domain related corpus provides better semantic representation of terms. Medical corpora can be classified into four categories.

2.1. Electronic Health Record (EHR)

In recent times, Electronic Health Records have become first option to store patient details in most of the hospitals [14]. EHRs include both structured data like diagnostic codes, procedure codes, medication codes, laboratory results etc. as well as unstructured data like clinical notes written by health professionals [15]. EHRs containing rich clinical information have become an invaluable source of data for many clinical informatics applications [16,17]. Some of the research studies have used publicly available EHR data while others have used private EHR data. MIMIC Dataset [18,19] is the largest publicly available EHR dataset and is described below.

2.1.1. MIMIC Dataset

Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) [18,19] is a publicly available ICU dataset developed by MIT Lab. It includes demographics, vital signs, laboratory tests, medications, and more. MIMIC-II [18] contains data collected from Intensive Care Units of Beth Israel Deaconess Medical Center from 2001 to 2008 while MIMIC III [19] consists of data collected in between 2001 and 2012 from the same medical center. The data in MIMIC datasets is deidentified and can be used for research purpose. But prior to access, agreement to data use and completion of a training course is mandatory.

2.2. Medical related Social Media Corpus

In recent times, social media evolved as a medium of expression for internet users. Medical related social media corpus includes tweets posted by individuals, questions and answers in online discussion forums related to health issues. In Twitter[1], users express health related concerns in short text of 140 characters while health discussion forums consists of health related questions raised and the corresponding answers. Some of the popular health discussion forums are MedHelp[2], DailyStrength[3], AskAPatient[4] and WebMD[5]. Social media text is highly informal and conversational in nature with lot of misspelled words, irregular grammar, non-standard abbreviations and slang words. Moreover, users describe their experiences in non-standard and descriptive words. Analysis of medical social media text which contains rich medical information can provide new medical insights and improved health care.

2.3. Online medical knowledge sources

Online medical knowledge sources contain medicine and health related information which is created and maintained by medical professionals. Merriam-Webster Medical Thesaurus[6], Merriam-Webster Medical Dictionary[7] and Merck Manual[8] are some of the online medical knowledge sources. Merriam-Webster Medical Thesaurus consists of word definition along with example sentence, synonyms, related words and antonyms while MerriamWebster Medical Dictionary consists of word definition along with multiple example sentences and synonyms. Merck Manual is a medical text book having articles related to various topics including disorders, drugs and tests. From these sources, corpus can be built and adopted by any embedding model to generate embeddings. eMedicine[9] is an online website which consists of almost 6,800 (by December 2018) articles related to various topics in medicine like Emergency medicine, Internal medicine etc. Each article is authored by a certified specialist in the concerned area which undergoes four levels of peer view which includes review by Doctor of Pharmacy. Medical Subject Headings (MeSH)[10] is created and maintained by United States National Library of Medicine[11]. It is a controlled vocabulary used for indexing articles in PubMed and classifying diseases in clinicaltrials.gov

MedlinePlus[12] maintained by United States National Library of Medicine offers reliable and updated information on various topics related to health in an easy to understand language. It is a medical encyclopedia that has information over 1000 diseases and conditions. Sciencedaily[13] and Medscape[14] are two other online sources that provides latest news related to medicine.

2.4. Scientific literature

PubMed[15] maintained by United States National Library of Medicine, is a search engine for citations and abstracts of research articles published in the areas of life sciences and biomedicine. As of December 2018, PubMed has 14.2 million articles with links to full-text. Apart from this, it provides access to books with full text available. PubMed Central (PMC)[16] is a digital repository of research papers published in the areas of biomedicine and life sciences and it provides free access. As of December 2018, it has over 5.2 million articles. Table 1 gives a comparison of various medical corpora.

3. Medical codes

The primary motive behind EHR [20] is to record the patient information right from admission to discharge in a systematic way. Several classification schemes are available for recording relevant clinical information. For example, ICD (International Statistical Classification of

[1] https://twitter.com
[2] https://www.medhelp.org
[3] https://www.dailystrength.org/
[4] https://www.askapatient.com/
[5] https://www.webmd.com/
[6] https://www.merriam-webster.com/thesaurus
[7] https://www.merriam-webster.com/medical
[8] https://www.msdmanuals.com/
[9] https://emedicine.medscape.com/
[10] https://www.ncbi.nlm.nih.gov/mesh
[11] https://www.nlm.nih.gov/
[12] https://medlineplus.gov/
[13] https://www.sciencedaily.com/
[14] https://www.medscape.com/
[15] https://www.ncbi.nlm.nih.gov/pubmed/
[16] https://www.ncbi.nlm.nih.gov/pmc/

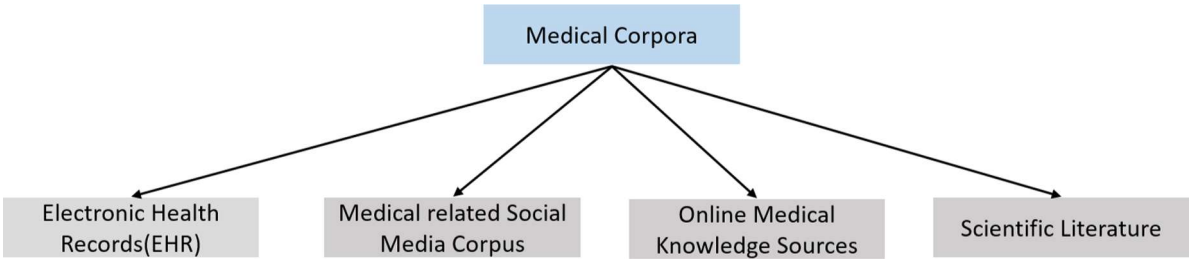

Fig. 1. Classification of medical corpora.

Diseases and Related Health Problems) - diagnosis codes, CPT (Current Procedural Terminology) - procedure codes, LOINC (Logical Observation Identifiers Names and Codes) - laboratory codes, and RxNorm-medication codes. Table 2 gives brief summary of various medical codes.

These standard codes are used to ensure consistency in recording patient information and other applications like reimbursement claims. Most of the US health care payment systems are based on these standard codes. As an example, the health care insurance companies pay reimbursements based on the medical codes assigned to clinical reports [21].

4. Embedding models

This section gives a brief overview as well as comparison of various embedding models like word2vec (Section 4.1), paragraph2vec (Section 4.2), glove (Section 4.3), fasttext (Section 4.4) and and Elmo (Section 4.5). Table 3 gives a summary of various embedding models and Table 5 gives a comparison of various embedding models.

Embedding is one of the promising applications of unsupervised learning as well as transfer learning because embeddings are induced from large unlabeled corpora and the prior knowledge captured in embeddings can be transferred to downstream tasks involving small datasets. Embedding models can be classified into Prediction based and Count based [22]. Prediction based models learn embeddings by predicting target word based on context words or vice versa. Count based models learn embeddings by leveraging global information such as word context co-occurrences in a corpus.

Research in learning distributed vector representations started with [1]. Several research studies [2–8] laid foundation for research in embeddings. Bengio et al. [4] proposed a neural network based model for the task of next word prediction. The model consists of hidden layer with *tanh* activation and output layer with *softmax* activation. The output layer produces probability of all the words in vocabulary for the given n-1 input words. In doing so, the model learns distributed representations of the words. The model overcomes the problems of curse of dimensionality and unseen sentences.

Collobert and Weston [7] were the first to demonstrate the usage of pre trained word embeddings. They proposed CNN based model which takes a sentence as input and outputs parts of speech tags, named entity tags, chunks etc. Finally, the model proposed by Mikolov et al. [9] gained lot of attention, brought immense popularity and made embeddings the first choice for text representation. Later models like Glove [10], FastText [11] were proposed. As these models are context unaware, they assign a single representation to a word ignoring its context which limits the quality of vector representation. To model context into word representations, models like ELMo [12], BERT [13] were proposed. ELMo is a feature based method i.e., ELMo vectors are used as input features in downstream tasks. BERT is a fine-tuning based method i.e., task specific layers are added and the model is fine-tuned using task specific labeled dataset.

4.1. Word2Vec

Inspired by distributional hypothesis [24,25] and neural language models, Mikolov et al. [9] proposed word2vec, a simple and efficient algorithm for inferring dense vector representations of words from a large unlabeled corpus. It is a shallow neural network model that learns word representations by optimizing objective function which involves both target word and context word. Word2vec builds vocabulary out of corpus and learns word representations by training a three layered neural network. Word2vec offers two models namely Continuous Bag of Words (CBOW) and Skip-gram. CBOW learns representations by predicting target word based on its context words while skipgram learns representations by predicting each of the context words based on target word. So, one has to choose one of the architectures and set values for hyper parameters like embedding size, context size, minimum frequency for a word to be included in vocabulary to generate word embeddings from a large corpus of unlabeled text. Table 4 gives a summary of various hyper parameters in word2vec model.

In Word2vec, there are two options to evaluate generated representations namely, distance and analogy. Distance option allow to retrieve the most semantically similar words for a given word in corpus along with cosine similarity score. In this context, cosine similarity score represents the degree of semantic similarity between the two words. Analogy option allows to find the linguistic regularities like 'king-man + women = ?'

4.1.1. Continuous Bag of Words (CBOW)

CBOW model learns embeddings by predicting the target word against its context words. CBOW model can be viewed as supervised model with context words as input and target word as output.

> For example, consider the sentence, "the black pen is on the red table". With a context window of size 2, (context, target) word pairs are ([black, pen], the), ([the, pen, is], black), ([the, black, is, on], pen) and so on.

As in Fig. 2, CBOW model consists of three layers namely input layer, hidden layer and output layer. The layers are connected by two weight matrices W and W'. The input layer takes one hot vectors of context words as input and output layer by applying softmax function predicts the one hot vector of target word. Error between the original and predicted vectors is back propagated to update the weight matrices W as well as W'. Finally for each word in vocabulary of given corpus, two vectors V_c and V_w are obtained (i.e., V_c is from W and V_w is from W'). The objective function of CBOW model is

$$J = \frac{1}{V}\sum_{i=1}^{V} logp\left(w_i \middle| w_{i-n}, \ldots w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}, \ldots, w_{i+n}\right) \quad (1)$$

where V is the size of vocabulary and n is the window size.

4.1.2. Skipgram

As in Fig. 2, Skip-gram model works exactly opposite to CBOW and

Table 1
Comparison of various medical corpora.

S.No.	Medical Corpus	Contains	Language used	Noisy	Access	Authors	Example
1	Electronic Health Records (EHR)	Patient information in the form of medical codes, laboratory results and clinical notes	Professional	Yes, with lot of unstandardized abbreviations and misspelled words in clinical notes	Restricted access because of sensitive information	Trained medical professionals in hospitals	MIMIC II [18] and MIMIC III [19]
2	Medical Social Media	Views and opinions related to health in the form of tweets. Health related questions and answers in Online Health discussion forums	Colloquial	Yes, with irregular grammar, slang words and misspelled words	Free Access	Common public	Twitter, AskAPatient, WebMD, MedHelp and DailyStrength
3	Online Medical Knowledge Sources	Medical words definitions, synonyms and related words, Medical articles and latest Medical news	Professional	No	Free Access	Trained medical professionals	Merriam Webster dictionary and Thesaurus, eMedicine, Medscape, MedLinePlus, Sciencedaily, MeSH and Merck Manuals
4	Scientific Literature	Abstracts, citations as well as full text of life sciences and biomedical research articles	Professional	No	Free Access	Researchers in life sciences and biomedical sciences	PubMed and PubMed Central

learns embeddings by predicting context words against target word. Skipgram model can be viewed as supervised model with target word as input and context words as output.

> For example, consider the sentence, “the black pen is on the red table”. With a context window of size 2, the (target word, context word) pairs are (the, [black, pen]), (black, [the, pen, is]), (pen, [the, black, is, on]) and so on.

The objective function of skipgram model is

$$J = \frac{1}{V}\sum_{i=1}^{V}\sum_{-n\leqslant j\leqslant n,\neq 0} logp\left(w_{i+j}\middle|w_j\right) \tag{2}$$

where V is the size of vocabulary and n is window size.

Each of CBOW and Skipgram models have their own advantages and disadvantages. In both the models, learning output vectors is computationally expensive. To address this issue, two methods namely Negative sampling and Hierarchical softmax is proposed. Negative sampling limits the number of output vectors to be updated while Hierarchical softmax is based on Huffman tree. Negative sampling works well with in-frequent words, whereas Hierarchical softmax works well with frequent words [26].

4.2. Doc2Vec

Paragraph2vector popularly known as Doc2vec, is an extension to Word2vec [9] and is proposed by Le et al. [23]. It is an unsupervised model which maps variable length text like sentences, paragraphs and documents to dense vector representations. Doc2vec learns dense vectors representations for both variable length text and words in the corpus. It offers two models namely Distributed Bag of Words (DBOW) and Distributed Memory (DM).

4.2.1. Distributed Memory (DM)

DM model is similar to Continuous Bag of Words model of Word2vec [9]. CBOW predicts the center word from context words while DM predicts next word using the concatenation or average of the vectors of paragraph and context words. As in Fig. 3, it consists of three layers. First layer takes the vectors of paragraph and context words as input. Second layer concatenates or average both these vectors. The final layer which is a classifier, predicts the vector for the next word.

4.2.2. Distributed Bag of Words (DBOW)

DBOW model is similar to Skipgram model of Word2vec. Skipgram predicts context words from center word while DBOW predicts context words using paragraph. As in Fig. 4, it consists of three layers. First layer takes the vector of paragraph as input. Second layer is the hidden layer. The final layer which is a classifier, predicts the vectors of context words. In both DBOW and DM models, the matrix D has dense vector representations. Each column of D is an embedding of the variable length text. Compared to DM model, DBOW model is simple, needs less memory. DM model stores softmax weights as well as word vectors while DBOW model stores only softmax weights. According to Le and Mikolov [23] DM works well for most of the tasks but recommended to use a combination of vectors from DM and DBOW, as the combination gives consistent results across tasks.

4.3. Glove

Global Vectors for word representations popularly knowns as “Glove” [10] is proposed by Pennington et al. Methods like LSA [6] which uses matrix factorization utilize global co-occurrence statistics but perform poorly in word analogy task. While methods like Word2vec do well in word analogy tasks but poorly utilize global cooccurrence

Table 2
Summary of various medical codes.

Schema	Description	Number of codes	Examples
ICD-10 (Diagnosis)	Prepared by World Health Organization(WHO) and contains codes for disease, signs and symptoms etc.	68,000	'R070': Pain in Throat 'H612': Impacted cerumen
CPT (Procedures)	Prepared by American Medical Association(AMA) and contain codes for medical, surgical and diagnostic services	9641	'90658':Flue Shot '90716': Chicken Pox Vaccine
LOINC (Laboratory)	Prepared by Regenstrief Institute, a US nonprofit medical research organization and contain codes for laboratory observations	80,868	'8310-5': Body Temperature '5792-7': Glucose
RxNorm (Medications)	Prepared by US National Library of Medicine and is a part of UMLS. It contains codes for all the medications available in US market.	1,16,075	'1191': Aspirin '215256': Anacin

Table 3
Summary of embedding models.

Model	Architecture	Advantages	Disadvantages
CBOW [9]	Log Bilinear	Faster compared to skipgram model. Represents frequent words well.	Ignore morphological information as well as polysemy nature of words No embeddings for OOV, misspelled and rare words.
Skipgram [9]	Log Bilinear	Efficient with small training datasets. Represents infrequent words well.	Ignore morphological information as well as polysemy nature of words No embeddings for OOV, misspelled and rare words.
PV-DM [23]	Log Bilinear	PV-DM alone give good results for many of the tasks.	Compared to PV-DBOW, requires more memory as it is needed to store Softmax weights and word vectors.
PV-DBOW [23]	Log Bilinear	Need to store only the word vectors and so requires less memory. Compared to PV-DM, it is simple and faster.	Need to be used along with PV-DM to give consistent results across tasks
Glove [10]	Log Bilinear	Combines advantages of word2vec model in learning representations based on context as well as matrix factorization methods in leveraging global co-occurrence statistics.	Ignore morphological information as well as polysemy nature of words No embeddings for OOV, misspelled and rare words.
FastText [11]	Log Bilinear	Encode morphological information in word vectors. Embeddings for OOV, misspelled and rare words. Pretrained word vectors for 157 languages.	Computationally intensive and memory requirements increases with the size of corpus. Ignore polysemy nature of words.
ELMo [12]	BiLSTM	Generate context dependent vector representations and hence account for polysemy nature of words Embeddings for OOV, misspelled and rare words.	Computationally intensive and hence requires more training time.

Table 4
Summary of hyper parameters in Word2Vec model.

Parameter	Default Value	Meaning
size	100	Dimension of vector
window	5	Size of context window
min_count	5	Minimum frequency of a word to be included in vocabulary
workers	3	Number of threads to train the model
sg	0	0 means CBOW model is used and 1 means skipgram is used.
hs	0	1 for hierarchical softmax 0 and 'negative' is non-zero means negative sampling is used
negative	5	0 means, no negative sampling >0 means negative sampling is applied and the value represents number of noise words to be used.

statistics. The Glove model combines the advantages of Word2vec model in learning representations based on context as well as matrix factorization methods in leveraging global co-occurrence statistics.

The model is trained using a weighted least squares objective function such that error between model predicted values and global count statistics from training corpus is minimized. The authors illustrated the importance of ratio of co-occurrence probabilities and proposed the base model as

$$F\left(u_i, u_j, v_k\right) = \frac{P_{ik}}{P_{jk}} \tag{3}$$

where u_i, v_j are focal word vectors and v_k is vector of context word. P_{ik} and P_{jk} represent the probability of words i and j to co-occur with word k.

To introduce linearity and avoid mixing vector dimensions, the authors introduced vector difference and dot product respectively.

$$F\left(dot\left(u_i - u_j\right), v_k\right) = \frac{P_{ik}}{P_{jk}} \tag{4}$$

Further to account for symmetry that word and context word are interchangeable in co-occurrence matrix, the model takes the form

$$u_i^T v_k + b_i + b_k = log\left(X_{ik}\right) \tag{5}$$

Here X_{ik} represents the co-occurrence frequency of word i with word k. Finally, the vectors are learned with weighted least squares objective function.

$$\sum_{i,k=1}^{V} f\left(X_{ik}\right)\left(u_i^T v_k + b_i + b_k - log\left(X_{ik}\right)\right)^2 \tag{6}$$

Here $u_i^T v_k + b_i + b_k$ represents model predicted values, $log\left(X_{ik}\right)$ represents value calculated from training corpus, V is vocabulary size. Further f(x) is a weighted function included in objective function so that rare or frequent co-occurrences are not over weighted and it is defined as

Table 5
Comparison of various embedding models.

Model	Type	Generate embeddings for	Vectors for OOV words	Encode morphological information	Use of global co-occurrence statistics	Word Vectors
CBOW	Prediction based	Words	No	No	No	Context Insensitive
Skipgram	Prediction based	Words	No	No	No	Context Insensitive
PV-DM	Prediction based	Sentences, Paragraphs and Documents	–	No	No	–
PV-DBOW	Prediction based	Sentences, Paragraphs and Documents	–	No	No	–
Glove	Count based	words	No	No	Yes	Context Insensitive
FastText	Prediction based	char n-grams	Sum of char n-grams	Yes	No	Context Insensitive
ELMo	Prediction based	words	Generated over character embeddings using CNN	Yes	No	Context Sensitive

$$f(x) = \left(\frac{x}{x_{max}}\right)^{\alpha}, \; if x \leqslant x_{max}$$

$$= 1, \; otherwise$$

The authors used stochastic gradient descent, $x_{max} = 100$ and $\alpha = 0.75$ to train the model.

4.4. FastText

Models like word2vec, glove treat words as atomic units and assign vector representations. These models completely ignore sub-word information and suffer from OOV issue. To leverage sub-word information and provide vectors for rare and OOV words, Bojanowski et al. [11] modified skipgram model and proposed FastText embedding model. In this model, vectors are learned from character n-grams and word representation is obtained as sum of its character n-grams. FastText by constructing word representation using its n-grams vectors, leverages sub word information and also offer quality vectors for rare and OOV words as their n-grams appear in training corpus.

Let G represents the set of all character n-grams in the training corpus. For a word w, G_w represents set of its character n-grams and z_g be the vector representation of each $g \in G$. For example, for n = 3, $G_{book} = \{ < bo, boo, ook, ok >, < book > \}$. Here the vectors are learned by predicting context words which is expressed as a set of binary classification tasks. For each word, negative log-likelihood is calculated using context words and randomly sampled negatives as

$$\sum_{c \in C_t} log\left(1 + e^{-s(w_t, w_c)}\right) + \sum_{n \in N_{t,c}} log\left(1 + e^{s(w_t, n)}\right)$$

Sum of negative log-likelihood of all the words gives the objective function

$$\sum_{t=1}^{T} \left[\sum_{c \in C_t} l\left(s\left(w_t, w_c\right)\right) + \sum_{n \in N_{t,c}} l\left(-s\left(w_t, n\right)\right) \right]$$

Here T is vocabulary size, C_t is set of context words for word t, $N_{t,c}$ is set of negative samples for word t and context word c, l is logistic loss function defined by $l(x) = log(1 + e^{-x})$. Further s is scoring function which calculates similarity of two words and it is defined as

$$s\left(w, c\right) = \sum_{g \in G_w} z_g^T . v_c$$

where v_c is vector of context word c.

To reduce the memory requirements, the authors applied hashing function which maps each character n-gram to unique value. Except n-gram size, the other hyper parameters are same as in skipgram. The authors suggested to use n-gram size of 3 to 6.

4.5. ELMo

Traditional embedding models like Word2Vec, Glove, FastText assign a single vector representation to a word independent of the context in which it is used. However, meaning of a word changes according to the context in which it is used. In recent times, a number of models like CoVE [27], TagLM [28], Context2vec [29] have been proposed to generate context dependent representations. However these models have some drawbacks. CoVE needs labeled data to generate context dependent representations and use zero vectors for OOV words. Further CoVE, TagLM and Context2vec models make use of only last layer representations. Peters et al. [12] proposed ELMo (Fig. 5) which generates embeddings for a word considering its context, by making use of Character embeddings and BiLSTM. Unlike CoVE, TagLM and Context2Vec models, ELMo makes use of all the three layer vectors i.e., the final representation of a word is obtained as task specific weighted

Table 6
Summary of research works which applied character embeddings.

Research Work	Model	Character Embedding size	Task
[30]	CNN	300	Medical Concept Normalization
[31]	BiLSTM	32	Medical Concept Normalization

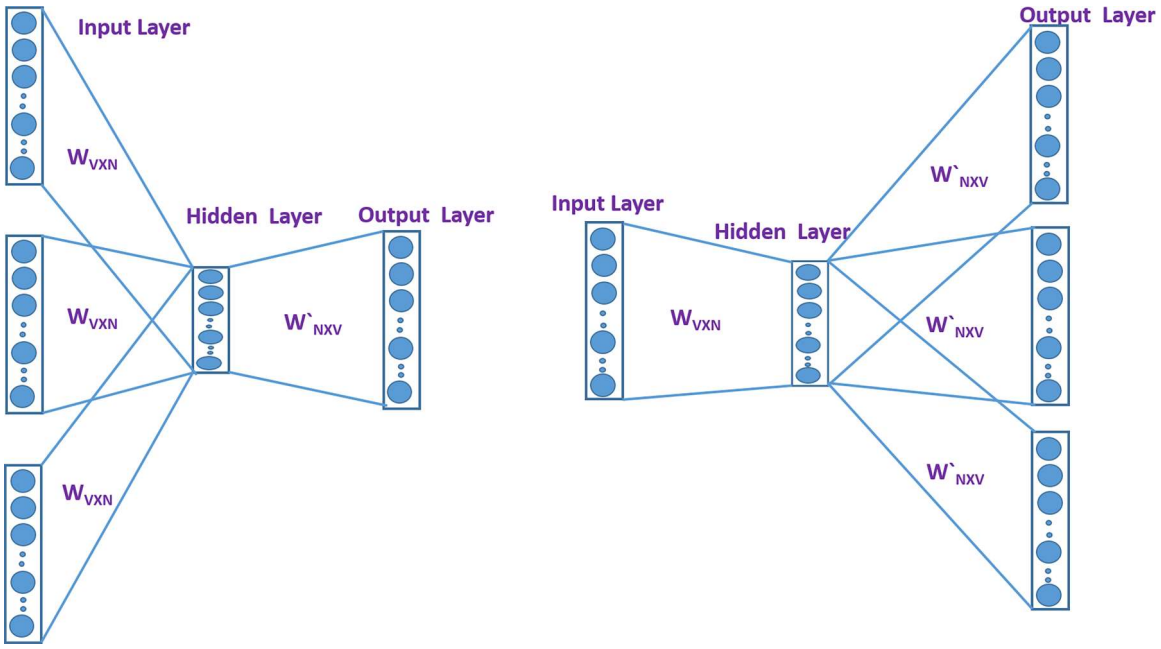

Fig. 2. CBOW and Skipgram models.

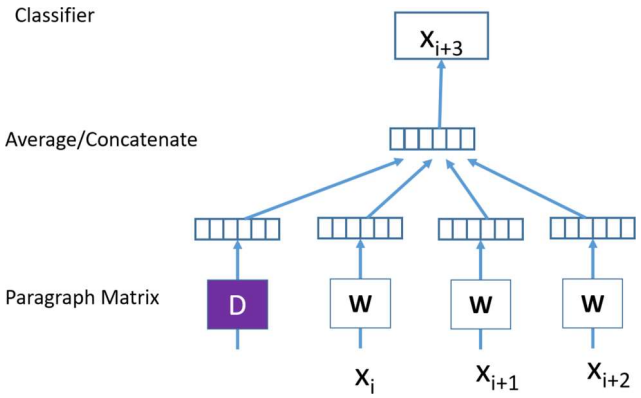

Fig. 3. DM Model.

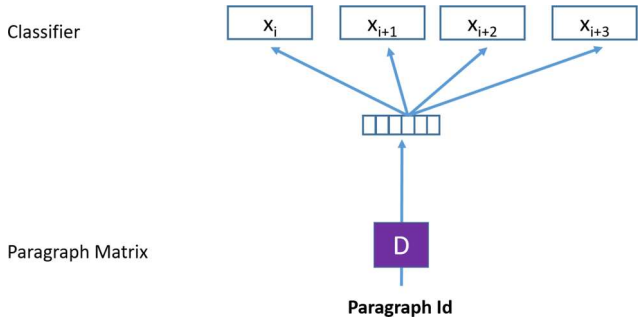

Fig. 4. DBOW Model.

average of all the three layer vectors. ELMo vectors are

- **deep** as they are obtained from three layer vectors
- **context sensitive** as they assign different representations to a word depending on its context
- **char based** as they generate representations of input words using char embeddings and CNN
- **versatile** as they can be used everywhere like traditional embeddings

CNN with character embeddings as input, generates context insensitive word vectors. With these word vectors as input, two layer BiLSTM is trained using language modeling objective.

$$\sum_{i=1}^{V} (\log p(t_i|t_1, t_2, ..,t_{i-1};\theta) + \log p(t_i|t_{i+1}, t_{i+2}, ..,t_V;\theta))$$

where θ represents parameters of CNN, BiLSTM and softmax layer.The first part represents forward language model which computes probability of given sequence of words $t_1, t_2, ..., t_V$ as

$$p\left(t_i \middle| t_1, t_2, ..,t_V\right) = \prod_{i=1}^{V} p\left(t_i \middle| t_1, t_2, ..,t_{i-1}\right)$$

The second part represents backward language model which process the given sequence in reverse and calculate probability as

$$p\left(t_i \middle| t_1, t_2, ..,t_V\right) = \prod_{i=1}^{V} p\left(t_i \middle| t_{i+1}, t_{i+2}, ..,t_V\right)$$

For each token t_i, ELMo model computes three vector representations $h_{i,0}^{LM}$, $h_{i,1}^{LM}$ and $h_{i,2}^{LM}$. $h_{i,0}^{LM}$ represents context insensitive vector generated by CNN with character embeddings as input, $h_{i,1}^{LM}$ is the hidden state vector obtained by the concatenation of forward and back hidden states ($\overrightarrow{h}_{i,1}^{LM}$ and $\overleftarrow{h}_{i,1}^{LM}$) of first BiLSTM layer, $h_{i,2}^{LM}$ is the hidden state vector obtained by the concatenation of forward and back hidden states ($\overrightarrow{h}_{i,2}^{LM}$

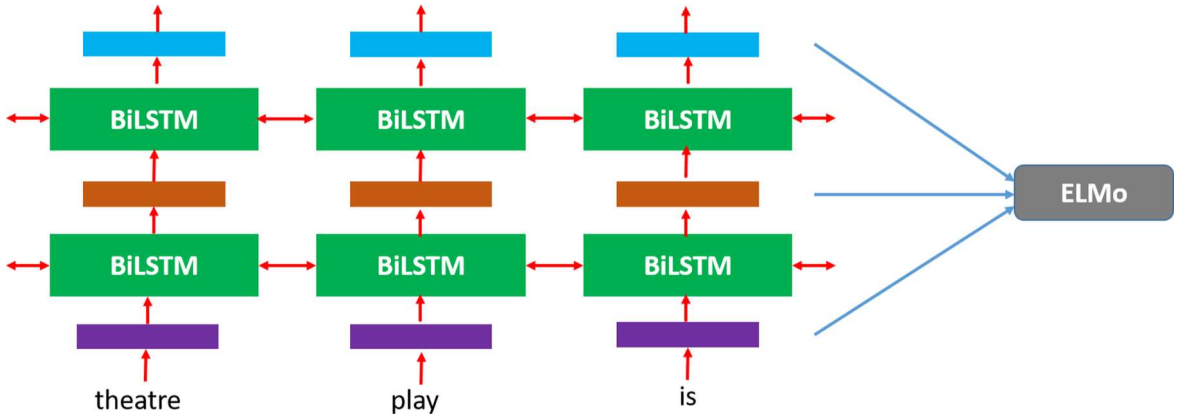

Fig. 5. ELMo model.

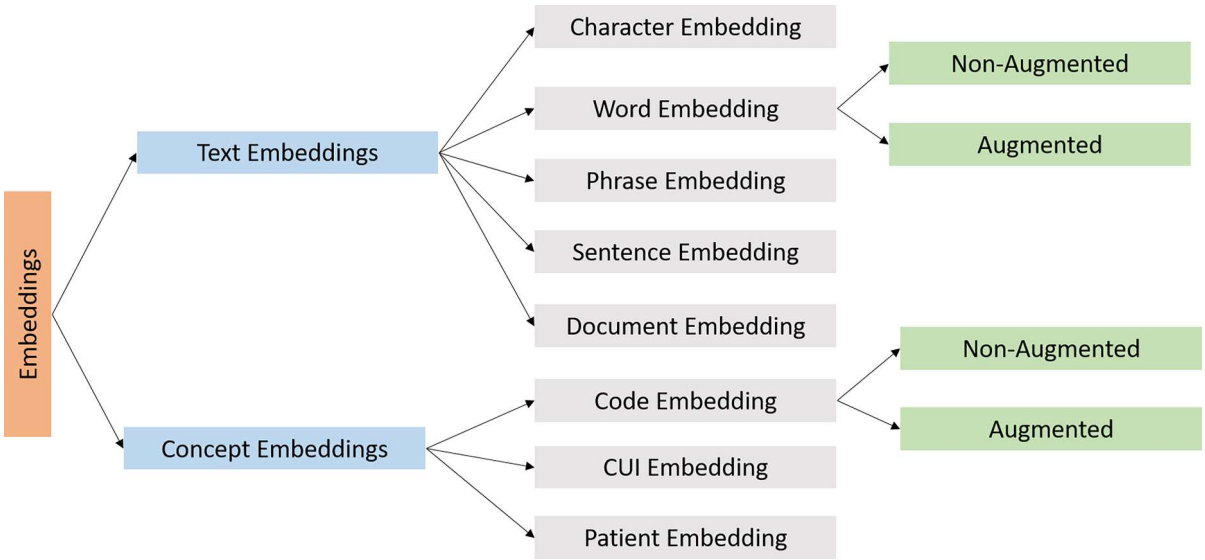

Fig. 6. Classification of clinical embeddings.

Table 7
Summary of various clinical embeddings.

Type of Embedding	Key Points	Papers
Char Embeddings	Include morphological information and can generate embeddings for OOV, misspelled and rare words. Meaning is not encoded	[30]
Word Embeddings	Incorporate syntactic and semantic information but not morphological information. No embeddings for OOV and misspelled words.	Non-Augmented [32–48,21,49–54,54–61] Augmented [62–69]
Code Embeddings	Finds applications in various tasks in health care analytics Ignoring sequential information in codes affects the quality of inferred embeddings.	Non-Augmented [70–73] Augmented [67,74]
CUI Embeddings	Encode domain knowledge from UMLS. Finds applications in information retrieval and analytics related to clinical domain.	[71,75]
Patient Embeddings	Finds applications in many tasks related to clinical domains. Ignoring sequential information in codes can potentially affect the quality of inferred embeddings.	[76–84]
Phrase Embeddings	Phrase embedding can be generated from the aggregation of embeddings of words in the phrases or directly using word2vec or paragraph2vec models.	[85,86]
Sentence Embeddings	Sentence embedding can be generated from the aggregation of embeddings of words in the sentence or directly using paragraph2vec model.	[87–89]
Document Embeddings	Document embedding can be generated from the aggregation of embeddings of sentences or directly using paragraph2vec model.	[90–93]

and $\overleftarrow{h}_{i,2}^{LM}$) of final BiLSTM layer. A task specific weighted average of these three vectors gives final representation of token t_i.

$$ELMo_i^{task} = r^{task} \sum_{j=0}^{2} s_j^{task} h_{ij}^{LM} \tag{7}$$

Here i and j are indices of word and BiLSTM layer, r^{task} represents task specific scaling factor, s_j^{task} represents softmax normalized weights.

5. Classification of medical embeddings

In this section, we discuss about various types of clinical embeddings. Depending on whether embedding map variable length text or concept, clinical embeddings can be broadly classified into two

Table 8
Summary of Clinical NLP tasks.

Task	Research Paper	Type of Embedding	Embedding Model	Corpus
Clinical Abbreviation Expansion	[3]	Word Embeddings	Word2Vec	Clinical Notes, Wikipedia Articles, ICU related Books and papers
MCN	[4]	Word Embeddings	Word2Vec	Merriam-Webster Thesaurus, Merriam-Webster Medical Dictionary, Clinical Text, Health Related Tweets
	[10]	Word Embeddings	Word2Vec	Health related Reviews, PubMed Literature
	[11]	Word Embeddings	Word2Vec	Google News Corpus, BioMed Literature
	[12]	Word Embeddings	Word2Vec	Google News Corpus, Generic Tweets and Drug Related Tweets
	[57]	Character, Word and Sentence Embeddings	Word2Vec and BiLSTM	Chinese Medical Corpus
Text Classification	[15]	Word Embeddings	Word2Vec	Two private EMR datasets and medical text book (the 7th edition of internal medicine)
	[16]	Word and CUI Embeddings	Word2Vec	MIMIC-III
	[19]	Word Embeddings	Word2Vec and Glove	Medical and Generic Corpus
	[22]	Word Embeddings	Word2Vec	Google News Corpus and Medical Corpus
	[24]	Word and Sentence Embeddings	Word2Vec and Doc2Vec	PubMed articles and Merck Manuals
	[46]	Word and Document Embeddings	Word2Vec and Doc2Vec	MIMIC-II and MIMIC-III
NER	[17]	Word Embeddings	Word2Vec	Health related Reviews, PubMed Literature
	[18]	Word Embeddings	Word2Vec	Drug related Reviews and Tweets
	[26]	Word Embeddings	Word2Vec	PubMed articles, Wikipedia articles and clinical notes
	[27]	Word Embeddings	Word2Vec	PubMed articles, Wikipedia articles and clinical notes
	[31]	Word Embeddings	Word2Vec and FastText	PubMed articles, Wikipedia articles and clinical notes
	[51]	Word Embeddings	Word2Vec	Health Related reviews
ICD Coding	[2]	Augmented Word Embeddings	Word2Vec	Google News Corpus, PubMed Literature and Wikipedia
	[20]	Word Embeddings	FastText	MIMIC-III
	[21]	Word Embeddings	Glove	MIMIC-III
	[23]	Word Embeddings	Word2Vec	Health Related Reviews, PubMed Literature, Wikipedia and Google News Corpus
	[28]	Word Embeddings	Glove	Common crawl
	[29]	Word Embeddings	Word2Vec	Google News Corpus
	[35]	Word Embeddings	Word2Vec	MIMIC-II and MIMIC-III
	[42]	Word Embeddings	Word2Vec	MIMIC-II and MIMIC-III
	[46]	Word and Document Embeddings	Word2Vec and Doc2Vec	MIMIC-II and MIMIC-III
IR	[25]	Word and Sentence Embeddings	Word2Vec, Glove and Doc2Vec	MEDLINE articles
Clinical Predictions	[8]	Code Embeddings	Word2Vec	Private EHR
	[30]	Patient Embeddings	Word2Vec	Private EHR
	[33]	Patient Embeddings	Word2Vec	Private EHR
	[47]	Code Embeddings	Word2Vec	Private EHR
	[53]	Patient Embeddings	Word2Vec	Private EHR
	[59]	Patient Embeddings	SDAE	Private EHR
Relation Classification	[32]	Word Embeddings	Word2Vec	Google News Corpus and MIMIC III
	[56]	Word Embeddings	Word2Vec	MIMIC-III
De-identification	[40]	Word Embeddings	Word2Vec, RNNLM	Google News Corpus and Broadcast News Corpus
	[49]	Word Embeddings	Glove	Wikipedia
Patient Similarity	[54]	Patient Embeddings	Word2Vec	Private EHR dataset

categories, namely Text embedding and Concept embedding. Text embedding maps variable length text like character, word, phrase, sentence or document to vectors while concept embedding maps medical code or UMLS Concept Unique Identifier (CUI) or patient information to vectors. As shown in Fig. 6, depending on the granularity of text they map to vectors, text embeddings can be further into five types and depending on the type of concept they map, concept emebddings can be further classified into three types. Depending on whether embedding is augmented or not with domain specific information, word as well as code embeddings can be further classified into Non-Augmented and Augmented types. Table 7 contains a summary of various clinical embeddings and Table 8 contains a summary of various clinical NLP tasks with embeddings as input features.

5.1. Character Embeddings

Character embedding models consider character as an atomic unit and maps it to a fixed length dense vector. Character level word representation is obtained from the embeddings of the constituting characters using CNN or BiLSTM. Let V be the vocabulary of characters which includes alphabets, digits, special characters, < unk > for unknown characters and < pad > for padding. E represents the embedding matrix such that each row is the embedding of a character and it is randomly initialized. For a given word, lookup is performed and stacking of embeddings of characters in the word gives S. With S as input, CNN or BiLSTM generates character level word representation.

Character level word vectors encode morphological information like prefix and suffix as well as orthographic information. They offer quality vectors for OOV as well as rare words. As character level word vectors are learned along with parameters of downstream model, they encode task as well as domain specific information. Table 6 shows summary of research works which applied character embeddings.

Table 9
Summary of publicly available word embeddings.

Name	Model	Corpus	Dimension
GoogleNewsVec [9]	Word2vec	Google News Corpus	300
$Glove_{wiki+giga}$ [10]	Glove	Wikipedia and Gigaword	50,100,150 and 200
$Glove_{cw\_uncased}$ [10]	Glove	Common Crawl	300
$Glove_{cw\_cased}$ [10]	Glove	Common Crawl	300
$Glove_{twitter}$ [10]	Glove	Tweets	300
$Word2vec_{PMC}$ [94]	Word2vec	PMC full text articles	200
$Word2vec_{PubMed+PMC}$ [94]	Word2vec	PMC full text articles and PubMed abstracts	200
$Word2vec_{PubMed}$ [94]	Word2vec	PubMed abstracts	200
$Word2vec_{PubMed+PMC+Wiki}$ [94]	Word2vec	PubMed abstracts, PMC full text articles and Wikipedia articles	200
$BioWordVec_{Intrinsic}$ [95]	FastText	PubMed and MeSH	200
$BioWordVec_{Extrinsic}$ [95]	FastText	PubMed and MeSH	200
BioWordVec [95]	FastText	PubMed and MIMIC-III	200
HealthVec [40]	word2vec	Health Reviews	200
DrugTweetsVec [41]	word2vec	Drug related tweets	150
TweetsVec [96]	word2vec	Tweets	400
PubMedVec [68]	AiTextML	PubMed abstracts	100
Drug2Vec [68]	word2vec	PubMed + DrugBank	420

5.2. Word Embeddings

Word embedding models map words to dense vector representations as well as capture syntactic and semantic information. Context independent word embedding models like word2vec [9], glove [10] and fasttext [11] assign single vector representation for a word ignoring the context in which it appears. Context dependent word embeddings models like Elmo [12], BERT [13] assign different representations for a word depending on the context in which it appears. It requires lot of computing resources and time to generate word embeddings. So, many of the research groups released their pretrained word embeddings. For example, Mikolov et al. [9] inferred embeddings using skipgram model and google news corpus. The released pretrained model consists of embeddings for around 3 million words and phrases. Table 9 shows the details of various publicly available pretrained word embeddings. The table includes both general and domain specific word embeddings.

Word embeddings are inferred from large unlabeled corpus using an embedding model. The embeddings learned capture only syntactic and semantic information from the training corpus. Such embeddings are called Non-augmented word embeddings as the embeddings are not augmented with information from any other sources. The quality of embeddings can be improved using knowledge from ontology or updating embeddings in downstream task. Such embeddings are called Augmented embeddings.

5.3. Non Augmented Word Embeddings

In non-augmented word embeddings, word embeddings are learned from large unlabeled text corpus and then used as input features in downstream tasks. Table 10 shows the summary of research works which applied non-augmented word embeddings.

Refs. [34,35,64,36] exploited general as well as domain specific word embeddings in the task of medical concept normalization. For example, Lee et al. [34] explored the use of word embeddings generated from various medical knowledge sources using word2vec. Medical Concept Normalization maps health condition expressed in lay terms to standard medical concepts and is treated as multi class classification problem. Evaluation on two standard datasets showed that RNN and CNN models trained using embeddings learned from combined clinical data sources outperformed the baselines.

Refs. [38,39,42,45] used word embeddings as input features for deep learning based text classification models. For example, Shen et al. [38] generated word embeddings using skipgram over the corpus consisting of medical records and medical text book. They generated word clusters using Hierarchical Agglomerative Clustering and then added cluster center vectors to the embeddings of each word. LSTM with word cluster embeddings outperformed the baselines with just word embeddings in short text classification.

Refs. [43,44,46,21,65,50] applied word embeddings in the task of assigning ICD codes. For example, Li et al. [43] formulated the problem of disease diagnosis prediction as multi class classification with ten categories. They generated embeddings with a dimension of 128 using FastText on the MIMIC III [19] corpus. Results demonstrated that CNN model with FastText embeddings trained on MIMIC III outperformed the baselines.

Refs. [40,41,47,97,69,55,98,56–59,61] used word embeddings as input features in deep learning based NER models. For example, Miftahutdinov et al. [40] proposed a CRF architecture with hand crafted features and HealthVec embeddings for the task of identifying disease and drug related expressions in user comments.

Refs. [52,60,53,54] exploited word embeddings in the task of patient data de-identification which is basically a sequence labeling problem. For example, Sheta et al. [52] applied Elman [99] and Jordan [100] RNN architectures in the task of patient data de-identification in clinical records. They experimented with different word embeddings like RNNLM [8] trained on Broad News Corpus and CBOW trained on Google news corpus.

5.4. Augmented word embeddings

Word embeddings are induced from unlabeled corpus. Embeddings inferred from large corpus encode more information compared to embeddings inferred from small corpus. However, it is difficult to get large corpus in clinical domain. Pretrained embeddings released along with popular embedding models like word2vec, glove were trained on generic corpus. In these generic pretrained models, embeddings are missing for many domain specific words and quality of embeddings for domain specific words is poor. These factors, limits the use of pretrained embeddings inferred from general corpus in domain tasks. In order to make up for the size of corpus and improve the quality of inferred domain specific embeddings, task specific as well as domain specific information can be added. Some of the possible ways to include domain specfic information is fine-tuning (updating) pre-trained off-the-shelf word embeddings[62,64,65,69] in a domain task or use of domain knowledge from ontologies like UMLS [63,66]. Table 11 shows the summary of research works which utilized augmented word embeddings.

5.5. Code embeddings

EHRs contain patient information in the form of free text as well as medical codes. Medical codes are used to ensure consistency in

Table 10
Summary of research works which applied non-augmented word embeddings. Here N represents no and Y represents yes.

Research work	Embedding Model	Corpus	Use of hand crafted features	Use of word embedding clusters	Task	Model
[34]	word2vec	Medical Text from various sources	N	N	Medical Concept Normalization	GRU
[35]	word2vec	Health forum reviews	Y	N	Medical Concept Normalization	BiGRU + Attention
[36]	word2vec	Google News Corpus, Tweets and Drug Tweets	N	N	Medical Concept Normalization	Ensemble of LR and Bi-GRU
[38]	word2vec	Medical Records and Medical Text Book	N	Y	Short Text Classification	LSTM
[39]	word2vec	MIMIC-III	Y	N	Disease Text Classification	CNN
[40]	Word2vec	Health forum reviews	Y	N	NER	CRF
[41]	Word2vec	Drug related tweets	N	N	NER	CRF
[43]	FastText	MIMIC III	N	N	Disease diagnosis prediction	CNN
[44]	Glove	MIMIC III	N	N	ICD Coding	CNN + Attention
[45]	word2vec	Google News Corpus, Biomedical literature	N	N	Text Classification	CNN
[46]	word2vec	PubMed	Y	N	ICD Coding	LSTM
[47]	Word2vec	PubMed, Clinical Notes and Wikipedia	N	N	NER	LSTM-CRF
[48]	Word2vec	PubMed, Clinical Notes and Wikipedia	N	N	Medical Event Detection	GRU
[21]	Glove	Common crawl	N	N	ICD Coding	LSTM
[49]	Word2vec	MIMIC III	N	N	Relation Classification	LSTM
[50]	Word2vec	MIMIC II and MIMIC III discharge summaries	N	N	ICD Coding	CNN + Attention
[52]	Word2vec	Google News Corpus	N	N	Patient Data De-identification	RNN
[53]	Glove	Wikipedia and Gigaword	N	N	Patient Data De-identification	LSTM
[97]	Word2vec	Health forum reviews	N	N	ADR Extraction	Bi-LSTM CRF
[54]	Glove	Wikipedia and Gigaword	Y	N	Patient Data De-identification	LSTM
[55]	Word2vec	E-cigarette corpus	N	N	NER	BiLSTM
[56]	Glove	Wikipedia and Gigaword	N	N	NER	BiLSTM CRF
[57]	Glove	Common Crawl and MIMIC-III	Y	Y	NER	BiLSTM CRF
[58]	Word2vec	General Tweets	N	N	NER	BiLSTM
[59]	ELMo	Medical related Wikipedia pages and MIMIC-III	N	N	NER	BiLSTM CRF
[61]	ELMo	1 billion Word Benchmark	Y	N	NER	LSTM CRF

Table 11
Summary of research works which applied augmented word embeddings. Here fine tuning of pretrained embeddings means updating embeddings during training of downstream model.

Research Work	Embedding Model	Corpus	Source of Augmentation	Method of Augmentation	Intrinsic Evaluation	Extrinsic Evaluation
[62]	Word2vec	PubMed	Medical Claims Dataset and ICD-10	Fine Tuning of pretrained embeddings	–	ICD Coding
[63]	Word2vecf	MIMIC III	UMLS	Generated embeddings using (w, CUI) pairs	Word Relatedness	–
[64]	Word2vec	Google News Corpus	Task specific Labeled Dataset	Fine tuning of pretrained embeddings	–	Medical Concept Normalization
[65]	Word2vec	Google News Corpus	Task specific Labeled Dataset	Fine tuning of pretrained embeddings	–	ICD Coding
[66]	Word2vec	PubMed and Medical related Wikipedia	UMLS	Graph Regularization	Word Similarity and Relatedness	Biomedical IR
[68]	All-in-Text	PubMed	Labels of PubMed documents	Joint learning of embeddings for words and labels	Word Similarity and Relatedness	–
[69]	Word2vec	MEDLINE and Wikipedia	Task specific Labeled Dataset	Fine tuning of pretrained embeddings	–	NER

Table 12
Summary of publicly available code embeddings.

Name	Embedding Model	Corpus	Dimension
$Word2vec_{claims\_codes\_300}$ [71]	word2vec	Medical Claims Dataset	300

recording patient information. Dense vector representations of medical codes finds applications in various tasks in health care analytics. For example, diagnosis codes, procedure codes, laboratory codes and drug codes can be embedded into separate spaces. However, these concepts have a relation among them. For example, aspirin with the drug code '1191' is used to cure fever with the diagnosis code 'R50.80'. Having a separate embedding space results in difficulty in finding relation between various medical codes. Advantage of having a combined vector space of all the medical codes is that, a medication to a disease can be found by finding the nearest neighbors to the specific disease code [71]. Table 12 shows the summary of publicly available code embeddings.

Here training corpus is collection of patient visits. In each patient visit, health condition is recorded in terms of medical codes. So, a patient visit can be expressed as sequence of medical codes. The semantic information of a medical code can be inferred from others codes in the same visit. In skip gram model (Fig. 7), each focal code is used to predict context codes. In CBOW model (Fig. 8), context codes are used to predict focal codes.

Depending on whether embeddings are augmented or not with knowledge from external sources, code embeddings can be classified into two types namely, Non-augmented and Augmented code embeddings.

5.6. Non-augmented code embeddings

In non-augmented code embeddings, code embeddings are inferred using any of the embeddings models and then evaluated using intrinsic tasks like similarity and relatedness [71] or used as input features in downstream tasks like early detection of heart failure [70] and prediction of diabetes, congestive failure [73]. Table 13 shows the summary of research works based on non-augmented code embeddings.

5.7. Augmented code embeddings

To improve the quality of code embeddings, knowledge from medical ontologies can be utilized. Table 14 shows the summary of research works based on augmented code embeddings.

5.8. CUI embeddings

UMLS consists of Metathesaurus, Semantic Network, Specialist Lexicon and a set of software tools like Metamap. Metathesaurus consists of over 5 million concept names incorporated from 100 controlled vocabularies like SNOMED CT, ICD, Rx-Norm etc. It clusters all the synonym concepts from different vocabularies into a single concept and assigns a unique identifier called Concept Unique Identifier. CUI groups synonym concept names in different vocabularies and thus it acts as a mapping structure between controlled vocabularies. CUI consists of 8 characters starting with C followed by 7 digits. For example, CUI of head pain is C0018681. Mapping CUIs to dense vector representations and then use of CUI embeddings as input features, enhances the performance of downstream tasks with knowledge captured from UMLS. Table 15 shows the summary of publicly available CUI embeddings and Fig. 9 shows the generation of CUI embeddings using skipgram model.

De Vine et al. [75] presented concept based Skip-gram model to map UMLS CUIs to dense vector representations. In contrast to traditional skip-gram which learns embeddings for words, this model learns

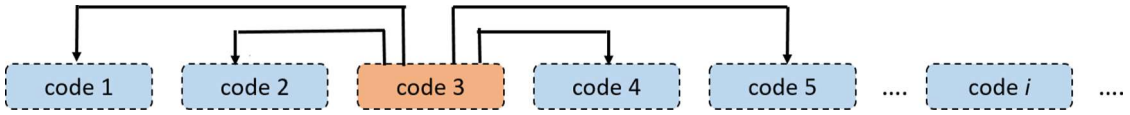

Fig. 7. Code embeddings using skipgram.

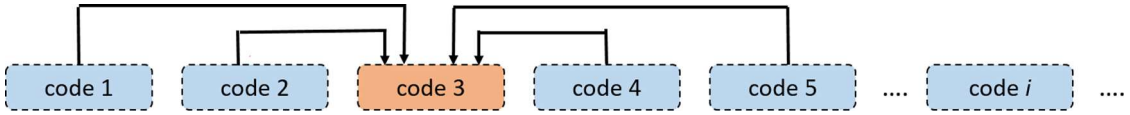

Fig. 8. Code embeddings using cbow.

embeddings for UMLS concepts extracted from clinical records and medical journal abstracts. Using MetaMap v11.2 [102], converted free text into sequence of concepts and then applied skip gram model. Evaluation on two medical word similarity datasets showed the effectiveness of UMLS CUI embeddings. This model finds applications in Information Retrieval and Analytics related to clinical domain. Unlike De Vine et al. [75], Choi et al. [71] learned representation of UMLS concepts using co-occurrence counts of concepts with in fixed time intervals derived from clinical narratives [103]. These outperformed the embeddings learned by [75] in the tasks of medical relatedness and similarity.

5.9. Patient embeddings

Patient information right from admission to discharge is recorded in EHR using free text as well as medical codes. Dense vector representation of patients is required in applications like prediction of clinical events and next visit time [78], prediction of unplanned readmission [77] etc. Patient embeddings encode patient information recorded in free text [81] or medical codes [76,78,80,82,83] or both [79] into dense fixed length vectors. Table 16 shows the summary of research works based on patient embeddings.

5.10. Phrase embeddings

Phrase embeddings maps phrases to fixed length dense vectors. Phrase embeddings can be generated from aggregation of embeddings of words in phrases or directly using paragraph2vec model.

Limsopatham and Collier [85] experimented with phrase based MT for the task of medical concept normalization. Initially, similarity score is computed between twitter phrase and the description of medical concept and then cosine similarity is calculated between twitter phrase and the description. Finally, twitter phrase is mapped to medical concept based on the linear combination of the two similarity scores.

Henry et al. [86] studied various dimensionality reduction techniques like skipgram, cbow, svd, explicit co-occurrence vectors and various multi term aggregation methods like sum, average, direct construction using compoundify tool or Meta Map in the context of semantic relatedness. Evaluation showed that i) none of the multi term aggregation method is better than the other which gives the flexibility in choosing methods ii) cbow embeddings with a size of 200 outperformed others.

5.11. Sentence embeddings

Sentence embedding maps sentence to fixed length dense vectors. Sentence embedding can be generated from the aggregation of embeddings of words in the sentence or directly using paragraph2vec model. For example, Zhang et al. [88] generated sentence embeddings using paragraph2vec [23] over psychiatric notes from CEGS N-GRID 2016 Challenge [109], psychiatric forum data from WebMD and MIMIC II [18] incrementally. Luo et al. [89] to map disease and procedure names in Chinese Discharge Summaries to standard names, applied sentence embeddings. They considered the final hidden state vector of BiLSTM as sentence embedding and here the input to BiLSTM is word vectors generated using word2vec and Chinese medical corpus. Table 17 shows the summary of research works based on sentence embeddings.

5.12. Document embedding

Document embedding maps documents to fixed length dense vectors. Document embedding can be generated from the aggregation of embeddings of sentences or directly using paragraph2vec model. For example, Li et al. [92] generated document embeddings by training paragraph2vec model over MIMIC III discharge summaries. Table 18 shows the summary of research works which utilized document embeddings.

6. Evaluation of embeddings

The quality of embeddings can be assessed in two ways namely intrinsic and extrinsic. Intrinsic evaluation looks how well the induced embeddings are able to encode syntactic and semantic information. Some of the tasks used in intrinsic evaluation are Clustering [72], Nearest Neighbor Search (NNS) [72], Similarity and Relatedness [63,71,75,66,68]. Similarity and Relatedness are the most commonly used tasks in intrinsic evaluation. For example, the datasets used in word similarity or word relatedness evaluation tasks like UMNSRS Similarity [105], UMNSRS Relatedness [105], MayoSRS [106], Pedersen's dataset [107], Hliaoutakis's dataset [108] consists of word pairs along with a similarity or relatedness score assigned by medical experts. Correlation between the cosine similarity of word vectors and expert assigned score is a measure of the quality of embeddings [63,66,68]. Table 19 shows the statistics of various word similarity and relatedness datasets.

In extrinsic evaluation methods, embeddings are used as input features in downstream tasks like Named Entity Recognition, Medical Concept Normalization, Medical Text Classification etc. and improvement in the performance of model is a measure of quality of embeddings. Extrinsic evaluation is necessary to find the effectiveness of embeddings in real world tasks. Table 20 shows intrinsic and extrinsic evaluation tasks in various clinical embeddings.

7. Challenges and solutions

In this section, we discuss various challenges in embeddings and highlight possible solutions from surveyed research papers.

7.1. Small size of clinical corpus

Embeddings are induced from unlabeled corpus and size of corpus is one of the factors which influence quality of inferred embeddings i.e., embeddings induced from large corpus encode more information

Table 13
Summary of research which applied non-augmented code embeddings.

Research Work	Embedding Model	Corpus	Codes Embedded	Intrinsic Evaluation	Extrinsic Evaluation
[70]	Word2vec	Private EHR dataset	Diagnosis, Procedure and Medication	–	Early detection of heart failure
[71]	Word2vec	Private Medical Claims Dataset, Private clinical Records	Diagnosis, Procedure, Laboratory and Drug	Similarity and Relatedness, Nearest Neighbor Search	–
[72]	Word2vec + Attention	Private and Public EHR datasets	Diagnosis, Procedure, Laboratory and Drug	Clustering and Nearest Neighbor Search	–
[73]	Word2vec	Private EHR dataset	Diagnosis and Medication	–	Prediction of diabetes and congestive failure

Table 14
Summary of resarch works which applied augmented code embeddings.

Research Work	Embedding Model	Corpus	Source of Augmentation	Method of Augmentation	Intrinsic Evaluation	Extrinsic Evaluation
[67]	Glove	Private EHR and MIMIC-III	CCS Multi-level diagnosis hierarchy	Each code is represented as weighted sum of ancestors in Ontology using attention and then fine tuned	–	sequential diagnoses prediction and heart failure prediction.
[74]	Word2vec	Private EHR	ICD Medical Ontology	Each code is represented in terms of four fine grained codes from ICD Ontology	–	Prediction of total hospital cost and length of stay.

Table 15
Summary of publicly available CUI embeddings.

Name	Embedding Model	Corpus	Dimension
stanford_Word2vec$_{cuis_svd_300}$ [71]	word2vec	Clinical Notes	300
Word2vec$_{cui_200}$ [75]	word2vec	MedTrack + OHSUMED	200
Cui2Vec [101]	word2vec	Medical Claims + Clinical Notes + PMC	500
Word2vec$_{cui_100}$ [68]	AiTextML	PubMed abstracts	100

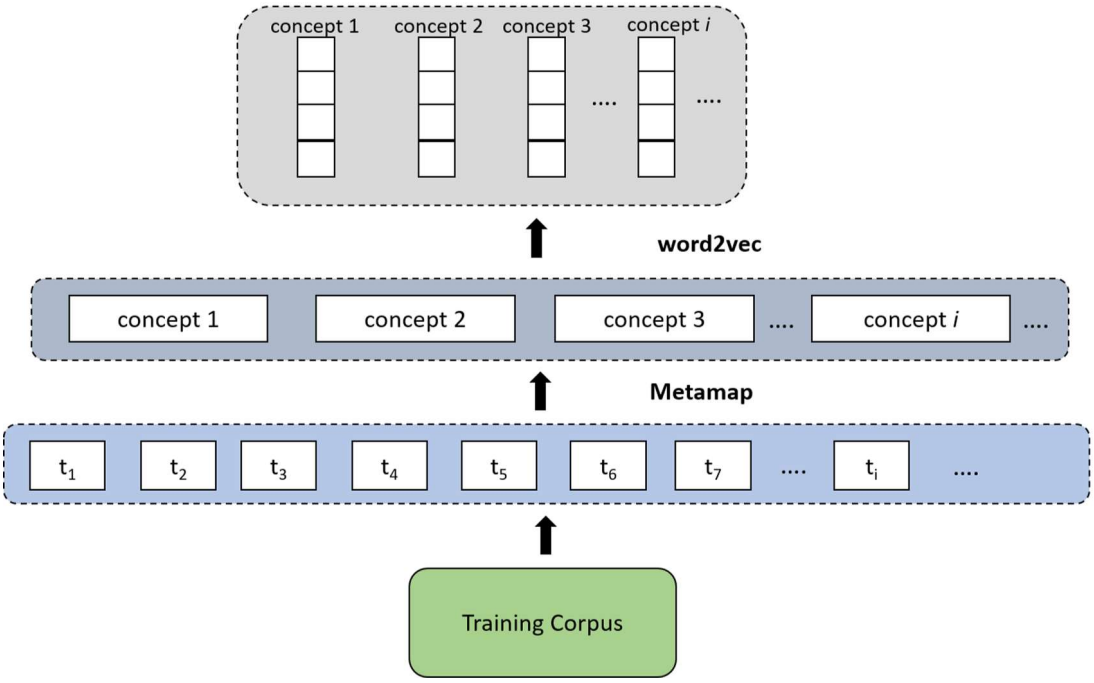

Fig. 9. Generation of CUI embeddings.

compared to embeddings induced from small corpus. In clinical domain, available corpora are small in size compared to general corpora. For example, Google News Corpus consists of around 100 billion tokens whereas MIMIC clinical notes consists of around 0.53 billion tokens, PubMed corpus consists of around 4.35 billion tokens. To make up for small size of clinical corpus, possible solutions are.

7.1.1. Inferring embeddings from combined corpora

Medical domain related text is available in various sources like Wikipedia, Medical Thesaurus, Medical Dictionary, PubMed and PMC, Clinical Notes from EHR. As the availability of large amount of text from a single source is not there, combining text from various sources and inferred embeddings from the combined corpora can potentially improve the quality of embeddings induced [34,75,47,48,50,88,59]. For example, Zhu et al. [59] generated domain specific ELMo embeddings using MIMIC III clinical notes and medical related Wikipedia articles, De Vine et al. [75] generated CUI embeddings using clinical notes and medical journal abstracts.

7.1.2. Use of knowledge from domain Ontologies

UMLS Metathesaurus clusters the concepts from over 100 controlled vocabularies like SNOMED CT, RxNorm and assigns all the concepts with same meaning, a CUI (Concept Unique Identifier). ICD contains disease and symptoms information, CPT contains procedures information, LOINIC contains laboratory observations information and RxNorm contains drugs related information. So, use of information from these medical ontologies which contain abundant domain knowledge can make up for size of corpora and improve the quality of embeddings inferred [63,75,39,66,67,74,68]. For example, Boag et al. [63] generated word embeddings using (w, CUI) pairs, Feng et al. [74] generated code embeddings using Private EHR and ICD ontology.

7.2. Multi sense embeddings

A word or a medical code can have more than one meaning referred to as multi sense. For example, a) the word 'bank' can refer place near river or a place where financial transactions are done, b) aspirin can be used a medicine for both fever and cardiovascular disease. However models like word2vec, glove, fasttext assign a single representation ignoring the multi sense nature which reduces the quality of inferred embeddings and affects the performance of model in downstream tasks. The possible solutions to generate multi sense embeddings are.

7.2.1. Contextualized representations

Models like ELMo [12], BERT [13] generate vector representations for a token depending on its context. These language representation models are trained over large volumes of text data using language modeling objective. ELMo representations are used as input features in downstream tasks while BERT can be used in both feature based and fine-tuning methods. In feature based method, the final hidden state vectors of BERT are used as input features while in fine-tuning method, task specific layers are added on the top of BERT and entire model is trained using task specific labeled data.

7.2.2. Use of topic information

[76,70,71] generated embeddings for medical codes without considering multi sense nature of codes i.e., assigned only one representation for codes with multi sense. Using context and topic

Table 16
Summary of research works which applied patient embeddings.

Research Work	Embedding Model	Corpus	Patient Embedding	Intrinsic Evaluation	Extrinsic Evaluation
[76]	Two Layer Neural Network	Private EHR	Output of second layer	–	Future code prediction and CRG prediction
[77]	Word2vec	Private EHR	Output of CNN with sequence of patient visits having medical codes as input	–	Prediction of unplanned readmission
[78]	Word2vec	Private EHR	Output of GRU with sequence of patient visits having medical codes as input	–	Prediction of medical codes and next visit time
[79]	Three layer neural network	MIMIC-III	Output of second layer	–	Comorbidity detection
[80]	Word2vec	Private EHR	Stacking of each visit vector representation obtained by addition of vectors of disease and procedure codes	–	Prediction of length of stay, total incurred charges and mortality rates
[81]	Paragraph2vec, Stacked Denoising Auto Encoder	MIMIC-III	Output of unsupervised model	–	Patient mortality prediction, Primary diagnostic and procedural category prediction and Gender prediction
[82]	Word2vec	Private EHR	Stacking of vectors of medical codes in patient visits	–	Patient Similarity
[84]	Stacked Denoising Auto Encoder	Private EHR	Output of final Denoising Auto Encoder	–	Disease Prediction

information, Qian et al. [112] learned multi sense embeddings for medical codes.

7.3. Domain adaptation of pretrained general embeddings

Pre-trained embeddings released with embeddings models like Word2vec, Glove, FastText, ELMo and BERT were trained on general corpus. As the semantic information encoded in vectors depends on genre of corpus, the utility of general pretrained embeddings in clinical tasks is limited. However, as these embeddings are inferred from large corpus, they encode lot of language information which is common across domains. To adapt pretrained embeddings to clinical domain, the possible options are.

7.3.1. Addition of task and domain specific knowledge

As embeddings are inferred over unlabeled text, embeddings capture both language information as well as semantic information. However, the quality of inferred embeddings can be improved further with the addition of domain knowledge. The possible ways are 1) while generating embeddings [63,66,67,74,68]. For example, Ling et al. [66] added knowledge from UMLS to word embeddings. They initially generated weighted graph and integrated it into word2vec model using graph regularization. The main drawback in this method of adding domain knowledge is, it is embedding model specific. 2) Fine-tuning embeddings in domain specific task [62,64,65,69]. For example, Patel et al. [62] improved pubmed embeddings by fine tuning them in ICD coding task. The main drawback in this method is, it requires labeled dataset. Fine tuning embeddings using a small labeled dataset can degrade the quality of embeddings. 3) fine-tuning embeddings using ontology [113–117]. This method of addition of knowledge overcomes the drawbacks in the above two methods. This method is neither embedding model specific nor it requires labeled dataset. Two of the popular models which fine tunes embeddings using ontology are Retrofitting [113] and Extrofitting [117]. However these methods are applicable only to context insensitive embeddings generated using word2vec, glove or FastText. So, there is a need to develop methods which can add ontology information to context sensitive embeddings also.

7.3.2. Ensemble of embeddings

Ensemble of word embeddings allows the downstream model to make use of different information encoded in embeddings induced from different corpora. Pre-trained word embeddings provide large coverage of vocabulary while domain specific word embeddings better represent terms. Further, genre of medical corpora effects the word embeddings induced. For example, health discussion forum embeddings better model colloquial medical terms while PubMed embeddings better model professional medical terms. So, using an ensemble of generic and domain specific word embeddings [36,45] or an ensemble of embeddings induced from different sources of medical corpora improves quality of embeddings [118].

7.3.3. Fine tuning and further pre-training

To adapt pretrained language model representations generated by ELMo [12] and BERT [13] to clinical domain, the possible options are 1) Further pre-train the model using domain specific corpus. For example, pretrained ELMo and BERT embeddings were inferred over Wikipedia and news crawl data, Wikipedia and Books corpus respectively. Instead of training these models from scratch using domain specific corpus, the pretrained models can be further trained using domain specific corpus like MIMIC III clinical notes or medical related Wikipedia pages. 2) Fine-tune the model using domain specific corpus. For example, pretrained BERT model is inferred over Wikipedia and Books corpus. The pretrained model can be adapted to clinical domain by adding task specific layers and fine tuning all the parameters using task specific labeled data. In fine tuning method, the model can be fine-

Table 17
Summary of research works which applied sentence embeddings.

Research Work	Embedding Model	Corpus	Sentence Embedding	Intrinsic Evaluation	Extrinsic Evaluation
[87]	Doc2vec	Merck Manual Articles	Output of Doc2vec	–	Medical Text Classification
[88]	Doc2vec	Psychiatric notes and MIMIC-III	Output of Doc2vec	–	Identification of psychiatric symptoms
[89]	Word2vec	Private EHR	Final hidden state vector of BiLSTM with word embeddings as input	–	Medical Concept Normalization

tuned using a single or multiple domain tasks. Fine-tuning approach is applicable only to BERT model but not ELMo model.

7.4. Sub-word information and OOV issue

It is necessary for word vectors to encode sub word information like prefix and suffix. However models like word2vec, glove consider word as an atomic unit and ignores sub-word information. As a result, the quality of vectors is limited and embeddings are missing for OOV words, rare and misspelled words. The possible solutions are 1) FastText [11] embeddings - This model learns vectors for character n-grams and represents a word as sum of its character n-gram vectors. By representing a word in terms of its character n-grams, sub-word information is leveraged and OOV, rare words can get quality vectors as their character n-grams appear in corpus. 2) Use of models like ELMo [12] and BERT [13] - ELMo model is char based so vectors encode sub-word information as well as eliminate OOV issue. In BERT, there will a fixed size of vocabulary consisting of all characters, most frequently occurring sub-words and words. OOV words are represented in terms of sub-words which eliminate OOV issue and leverage sub-word information.

7.5. Temporal Information

[37,67,71,75] explored various methods to learn concept embeddings (Code or CUI). In all these approaches, codes or CUIs are treated as words and embeddings are learned using others concepts within a fixed size window as contexts. In doing so, temporal information or temporal scope of medical concepts is ignored. This is because, two concepts with in a fixed size window need not be temporally close. Further, different concepts have different temporal scopes. For example, common cold persists for only days while diabetes persists for years [119]. So, learning embeddings using contexts in a fixed window cannot model temporal scope of medical concepts. Cai et al. [72] jointly learned embeddings and temporal scope of each medical concept. The proposed model uses attention mechanism to learn the temporal scope for medical concepts. We strongly believe this is an area which is to explored more.

8. Discussion

In this review paper, we provided a comprehensive survey of embeddings in clinical natural language processing and to our latest knowledge, it is the first attempt. We classified medical corpora into four types depending on the source, discussed each type in detail and finally provided a comparison to highlight characteristics of each corpus type. We discussed popular embedding models like word2vec, glove, fastText, doc2vec, ELMo and BERT and then compared them to highlight advantages and disadvantages of each. We broadly classified embeddings into two types depending on whether they map text or concepts. Further, text embeddings are classified into five types depending on the granularity of text they map and code embeddings are classified into three types depending on the concept (code, cui or patient) they map. We discussed various methods like intrinsic and extrinsic to evaluate embeddings and presented summary of evaluation tasks in various clinical embeddings. Finally, we discussed various challenges like *small size of clinical corpus*, *multi sense embeddings*, *domain adaptation of general embeddings*, *leveraging sub-word information*, *OOV issue*, *temporal information* and suggested possible solutions from the surveyed research articles. Even though embeddings became de facto standard for text representation and improved NLP models greatly, there are still some drawbacks.

First one is *interpretability*. Even though embeddings are widely used in NLP tasks, still they are opaque i.e., embeddings capture synatic and semantic information but it is not clear what exact properties are encoded. Factors like high dimensionality makes embeddings difficult to interpret. With poor interpretability, embeddings cannot be used in applications where reasoning for decisions is required. Most of research work focused on methods to generate embeddings. Only few research works like [120–127] in general domain and [128] in medical domain focused on interpretability. For example, [123,129] applied sparse coding to make embeddings interpretable, [120–122] proposed methods based on non-matrix factorization to make embeddings transparent. Chen et al. [128] evaluated word embeddings for various semantic relations using embeddings generated using word2vec and glove models over domain specific corpus. This is an area which is to be explore further to increase transparency of embeddings and hence increase their utility also.

Second one is *knowledge distillation*. BERT model advanced state-of-the-art performance in many NLP tasks. $BERT_{base}$ has 110 M parameters while $BERT_{large}$ has 340 M parameters. Both these models were trained on Wikipedia and Books corpus. Due to large number of parameters and large inference time, it is not feasible to implement these models in real world applications, particularly in applications with limited resources. Knowledge distillation proposed by [130,131] allows transfer of knowledge from large models like BERT into simpler models which are feasible to implement. In this way, satisfactory results can be obtained using simpler models itself without making any changes in model or use of additional data sets or features to train the model. This is a promising research direction which is still unexplored in clinical domain.

Third one is, *bias*. Bias refers to social or cultural inequality [132]. For example, when you think of 1) 'nurse', woman comes into mind and 2) 'doctor', man comes into mind which is essentially inequality. Embedding model which are trained on real word data which contains lot of bias, also exhibit the same kind of behaviour. For example, in embedding space, the word 'he' is closer to 'doctor' compared to 'she'. It is necessary to identify as well as remove bias in embeddings. Bolukbasi et al. [133] and Hoffman et al. [134] proposed methods to eliminate bias in general and clinical word embeddings generated using context insensitive embeddings. However, there is need to explore further in this topic to develop methods to eliminate bias in contextualized representations generated by models like ELMo and BERT in both general and clinical domain.

Fourth one is *evaluation of embeddings*. We have intrinsic and extrinsic methods to evaluate embeddings. But, sometimes there is no correlation between performance scores of embeddings in these methods. It is still not known which one is sufficient [135]. Faruqui et al. [136] discussed various problems in evaluating embeddings using similarity data sets. As reported in Table 20, most of the embeddings are evaluated intrinsically using similarity data sets. So, there is need to explore other intrinsic tasks as well as to develop data sets using which embeddings can be evaluated effectively.

Table 18
Summary of research works which applied document embeddings.

Research Work	Embedding Model	Corpus	Sentence Embedding	Intrinsic Evaluation	Extrinsic Evaluation
[104]	Doc2vec	MEDLINE articles	Output of Doc2vec	–	IR
[91]	Word2vec	MIMIC II and MIMIC III	Output of second GRU layer of Hierarchical Attention BiGRU model	–	ICD Coding
[92]	Doc2vec	MIMIC III Discharge Summaries	Output of Doc2vec	–	ICD Coding
[93]	Word2vec	Private Radiology Reports	Average of vectors of words in radiology report	–	Radiology report classification

Table 19
Statistics of various word relatedness and similarity data sets.

Dataset	Number of Word pairs	Type
UMNSRS Similarity [105]	566	Word Similarity
UMNSRS Relatedness [105]	588	Word Relatedness
MayoSRS [106]	101	Word Relatedness
Pedersen's dataset [107]	30	Word Similarity
Hliaoutakis's dataset [108]	34	Word Similarity

Table 20
Summary of intrinsic and extrinsic evaluation in clinical embeddings.

Type	Intrinsic Evaluation	Extrinsic Evaluation
Char	–	Medical Concept Normalization [30]
Word	Non-Augmented Similarity and Relatedness [109] Augmented Relatedness[63] Similarity and Relatedness [66,68]	Non-Augmented Clinical Abbreviation Expansion [33] Medical Concept Normalization [34,35,64,36] Medical Text Classification [38,42,45] Named Entity Recognition [40,41,47,48,110,97] ICD Coding [43,44,46,21,50,91] Relation Classification [49,111] Patient De-identification [52,60,53] Augmented Medical Coding [62] Biomedical IR [66] Medical Relation Classification [111]
Code	Non-Augmented Similarity and Relatedness [71] Clustering and Nearest Neighbor Search(NNS) [72]	Non-Augmented Heart Failure Detection [70] Risk Prediction [73] Augmented Sequence prediction task and Heart failure prediction task [67] Prediction of total hospital costs and length of stay (LOS) [74]
CUI	Similarity and Relatedness [71] Similarity [75]	–
Patient	–	Clinical Predictions [71] Unplanned Readmission [77] Medical Event Prediction [59] Prediction of clinical events and Next Visit time [78] Comorbidity detection task [79] Prediction of length of stay, total incurred charges and mortality rates [80]
Phrase	Similarity and Relatedness [86]	Medical Concept Normalization [85]
Sentence	–	Medical Text Classification [87] Identifying psychiatric symptoms [88] Medical Concept Normalization [89]
Document	–	ICD Coding [91,92] Radiology Report Classification [93]

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533.
[2] J.L. Elman, Distributed representations, simple recurrent networks, and grammatical structure, Mach. Learn. 7 (1991) 195–225.
[3] A.M. Glenberg, D.A. Robertson, Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning, J. Memory Lang. 43 (2000) 379–401.
[4] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.

[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[6] S.T. Dumais, Latent semantic analysis, Annu. Rev. Inf. Sci. Technol. 38 (2004) 188–230.

[7] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in, Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 160–167.

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

[9] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013, pp. 1–12.

[10] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[11] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.

[12] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2018, pp. 2227–2237. doi:https://doi.org/10.18653/v1/N18-1202.

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

[14] D. Charles, M. Gabriel, M.F. Furukawa, Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2012, ONC Data Brief 9 (2013) 1–9.

[15] G.S. Birkhead, M. Klompas, N.R. Shah, Uses of electronic health records for public health surveillance to advance public health, Annu. Rev. Public Health 36 (2015) 345–359.

[16] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary use of ehr: data quality issues and informatics opportunities, Summit Transl. Bioinformatics 2010 (2010) 1.

[17] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (2012) 395.

[18] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database, Critical Care Med. 39 (2011) 952.

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

[20] B. Shickel, P.J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, IEEE J. Biomed. Health Informatics 22 (2018) 1589–1604, https://doi.org/10.1109/JBHI.2017.2767063.

[21] S. Ayyar, O. Bear, Tagging patient notes with icd-9 codes, Proceedings of the 29th Conference on Neural Information Processing Systems, 2017.

[22] M. Baroni, G. Dinu, G. Kruszewski, Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2014, pp. 238–247.

[23] Q. Le, T. Mikolov, Distributed representations of sentences and documents, International Conference on Machine Learning, 2014, pp. 1188–1196.

[24] J.R. Firth, A synopsis of linguistic theory, 1930-1955, Studies in linguistic analysis, 1957.

[25] Z.S. Harris, Distributional structure, Word 10 (1954) 146–162.

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2, NIPS'13, 2013, pp. 3111–3119.

[27] B. McCann, J. Bradbury, C. Xiong, R. Socher, Learned in translation: Contextualized word vectors, in: Advances in Neural Information Processing Systems, 2017, pp. 6294–6305.

[28] M. Peters, W. Ammar, C. Bhagavatula, R. Power, Semi-supervised sequence tagging with bidirectional language models, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2017, pp. 1756–1765.

[29] O. Melamud, J. Goldberger, I. Dagan, context2vec: Learning generic context embedding with bidirectional lstm, in: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, 2016, pp. 51–61.

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

[31] S. Han, T. Tran, A. Rios, R. Kavuluru, Team uknlp: Detecting adrs, classifying medication intake messages, and normalizing adr mentions on twitter, in: SMM4H@ AMIA, 2017, pp. 49–53.

[32] J. Huang, K. Xu, V.V. Vydiswaran, Analyzing multiple medical corpora using word embedding, Proceedings of IEEE International Conference on Health Informatics (ICHI), 2016, pp. 527–533.

[33] Y. Liu, T. Ge, K. Mathews, H. Ji, D. McGuinness, Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion, in: Proceedings of BioNLP, 2015, pp. 92–97.

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

[35] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, V. Malykh, Medical concept normalization in social media posts with recurrent neural networks, J. Biomed. Inform. 84 (2018) 93–102, https://doi.org/10.1016/j.jbi.2018.06.006.

[36] M. Belousov, W.G. Dixon, G. Nenadic, Using an ensemble of linear and deep learning models in the smm4h 2017 medical concept normalisation task, SMM4H@ AMIA, 2017, pp. 54–58.

[37] J.A. Miñarro-Giménez, O. Marín-Alonso, M. Samwald, Exploring the application of deep learning techniques on medical text corpora, Stud. Health Technol. Informatics 205 (2014) 584–588.

[38] Y. Shen, Q. Zhang, J. Zhang, J. Huang, Y. Lu, K. Lei, Improving medical short text classification with semantic expansion using word-cluster embedding, Information Science and Applications, Springer, 2018, pp. 401–411.

[39] L. Yao, C. Mao, Y. Luo, Clinical text classification with rule-based features and knowledge-guided convolutional neural networks, Proceedings of IEEE International Conference on Healthcare Informatics Workshop (ICHI-W), 2018, pp. 70–71, , https://doi.org/10.1109/ICHI-W.2018.00024.

[40] 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. 155–166.

[41] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Am. Med. Inform. Assoc. 22 (2015) 671–681, https://doi.org/10.1093/jamia/ocu041.

[42] N. Pattisapu, M. Gupta, P. Kumaraguru, V. Varma, Medical persona classification in social media, Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, ASONAM '17, ACM, 2017, pp. 377–384, , https://doi.org/10.1145/3110025.3110114.

[43] C.Y. Li, D. Konomis, G. Neubig, P. Xie, C. Cheng, E.P. Xing, Convolutional neural networks for medical diagnosis from admission notes, CoRR abs/1712.02768, 2017.

[44] A. Karmakar, Classifying medical notes into standard disease codes using machine learning, CoRR abs/1802.00382, 2018. arXiv:1802.00382.

[45] N. Limsopatham, N. Collier, Modelling the combination of generic and target domain embeddings in a convolutional neural network for sentence classification, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, 2016, pp. 136–140. doi:https://doi.org/10.18653/v1/W16-2918.

[46] Z. Miftahutdinov, E. Tutubalina, Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks, 2017.

[47] A. Jagannatha, h. yu, Structured prediction models for rnn based sequence labeling in clinical text, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, , https://doi.org/10.18653/v1/D16-1082 pp. 856–865.

[48] A.N. Jagannatha, H. Yu, Bidirectional rnn for medical event detection in electronic health records, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2016, pp. 473–482. doi:https://doi.org/10.18653/v1/N16-1056.

[49] Y. Luo, Recurrent neural networks for classifying relations in clinical notes, J. Biomed. Inform. 72 (2017) 85–95, https://doi.org/10.1016/j.jbi.2017.07.006.

[50] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2018, pp. 1101–1111. doi:https://doi.org/10.18653/v1/N18-1100.

[51] Z. Jiang, L. Li, D. Huang, L. Jin, Training word embeddings for deep learning in biomedical text mining tasks, Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 625–628, , https://doi.org/10.1109/BIBM.2015.7359756.

[52] S. Yadav, A. Ekbal, S. Saha, P. Bhattacharyya, Deep learning architecture for patient data de-identification in clinical records, Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), The COLING 2016 Organizing Committee, 2016, pp. 32–41.

[53] F. Dernoncourt, J.Y. Lee, O. Uzuner, P. Szolovits, De-identification of patient notes with recurrent neural networks, J. Am. Med. Inform. Assoc. (2016), https://doi.org/10.1093/jamia/ocw156.

[54] J.Y. Lee, F. Dernoncourt, O. Uzuner, P. Szolovits, Feature-augmented neural networks for patient note de-identification, in: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 17–22.

[55] J. Xie, X. Liu, D. Dajun Zeng, Mining e-cigarette adverse events in social media using bi-lstm recurrent neural network with word embedding representation, J. Am. Med. Inform. Assoc. 25 (2017) 72–80.

[56] R. Chalapathy, E. Zare Borzeshi, M. Piccardi, Bidirectional lstm-crf for clinical concept extraction, Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), The COLING 2016 Organizing Committee, 2016, pp. 7–12.

[57] I.J. Unanue, E.Z. Borzeshi, M. Piccardi, Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition, J. Biomed. Inform. 76 (2017) 102–109.

[58] A. Cocos, A.G. Fiks, A.J. Masino, Deep learning for pharmacovigilance: recurrent

neural network architectures for labeling adverse drug reactions in twitter posts, J. Am. Med. Inform. Assoc. 24 (2017) 813–821.
[59] H. Zhu, I.C. Paschalidis, A. Tahmasebi, Clinical concept extraction with contextual word embedding, arXiv preprint arXiv:1810.10566, 2018.
[60] Y.-S. Zhao, K.-L. Zhang, H.-C. Ma, K. Li, Leveraging text skeleton for de-identification of electronic medical records, BMC Med. Inform. Decis. Mak. 18 (2018) 18, https://doi.org/10.1186/s12911-018-0598-6.
[61] Y. Tao, B. Godefroy, G. Genthial, C. Potts, Effective feature representation for clinical text concept extraction, arXiv preprint arXiv:1811.00070, 2018.
[62] K. Patel, D. Patel, M. Golakiya, P. Bhattacharyya, N. Birari, Adapting pre-trained word embeddings for use in medical coding, in: Proceedings of BioNLP, 2017, pp. 302–306.
[63] W. Boag, H. Kané, Awe-cm vectors: Augmenting word embeddings with a clinical metathesaurus, CoRR abs/1712.01460, 2017.
[64] N. Limsopatham, N. Collier, Normalising medical concepts in social media texts by learning semantic representation, in, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1014–1023.
[65] X. Zhang, R. Henao, Z. Gan, Y. Li, L. Carin, Multi-label learning from medical plain text with convolutional residual models, arXiv preprint arXiv:1801.05062, 2018.
[66] Y. Ling, Y. An, M. Liu, S.A. Hasan, Y. Fan, X. Hu, Integrating extra knowledge into word embedding models for biomedical nlp tasks, Proceedings of International Joint Conference on Neural Networks (IJCNN), 2017, pp. 968–975, , https://doi.org/10.1109/IJCNN.2017.7965957.
[67] E. Choi, M.T. Bahadori, L. Song, W.F. Stewart, J. Sun, Gram: Graph-based attention model for healthcare representation learning, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, ACM, New York, NY, USA, 2017, pp. 787–795. doi:https://doi.org/10.1145/3097983.3098126.
[68] E. Mencía, G. De Melo, J. Nam, Medical concept embeddings via labeled background corpora, in: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, European Language Resources Association (ELRA), 2016, pp. 4629–4636.
[69] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, H. Xu, Entity recognition from clinical texts via recurrent neural network, BMC Med. Inform. Decis. Mak. 17 (2017) 67, https://doi.org/10.1186/s12911-017-0468-7.
[70] E. Choi, J. Sun, A. Schuetz, W.F. Stewart, Using recurrent neural network models for early detection of heart failure onset, J. Am. Med. Inform. Assoc. 24 (2016) 361–370.
[71] Y. Choi, C.Y.-I. Chiu, D. Sontag, Learning low-dimensional representations of medical concepts, AMIA Summits Transl. Sci. Proc. (2016) 41.
[72] X. Cai, J. Gao, K.Y. Ngiam, B.C. Ooi, Y. Zhang, X. Yuan, Medical concept embedding with time-aware attention, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, https://doi.org/10.24963/ijcai.2018/554.
[73] Z. Che, Y. Cheng, Z. Sun, Y. Liu, Exploiting convolutional neural network for risk prediction with medical feature embedding, arXiv preprint arXiv:1701.07474, 2017.
[74] Y. Feng, X. Min, N. Chen, H. Chen, X. Xie, H. Wang, T. Chen, Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding, Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 770–777, , https://doi.org/10.1109/BIBM.2017.8217753.
[75] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, P. Bruza, Medical semantic similarity with a neural language model, Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, CIKM, 2014, pp. 1819–1822, , https://doi.org/10.1145/2661829.2661974.
[76] E. Choi, M.T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun, Multi-layer representation learning for medical concepts, Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, pp. 1495–1504, , https://doi.org/10.1145/2939672.2939823.
[77] P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, Deepr: A convolutional net for medical records, IEEE J. Biomed. Health Informatics 21 (2017) 22–30, https://doi.org/10.1109/JBHI.2016.2633963.
[78] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor ai: Predicting clinical events via recurrent neural networks, in: Proceedings of the 1st Machine Learning for Healthcare Conference, vol. 56, PMLR, 2016, pp. 301–318.
[79] D. Dligach, T. Miller, Learning patient representations from text, in: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, 2018, pp. 119–123. doi:https://doi.org/10.18653/v1/S18-2014.
[80] J. Stojanovic, D. Gligorijevic, V. Radosavljevic, N. Djuric, M. Grbovic, Z. Obradovic, Modeling healthcare quality via compact representations of electronic health records, IEEE/ACM Trans. Comput. Biol. Bioinf. 14 (2017) 545–554, https://doi.org/10.1109/TCBB.2016.2591523.
[81] M. Sushil, S. Šuster, K. Luyckx, W. Daelemans, Patient representation learning and interpretable evaluation using clinical notes, J. Biomed. Inform. 84 (2018) 103–113, https://doi.org/10.1016/j.jbi.2018.06.016.
[82] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, F. Wang, Measuring patient similarities via a deep architecture with medical concept embedding, Proceedings of IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 749–758, , https://doi.org/10.1109/ICDM.2016.0086.
[83] S. Dubois, N. Romano, D.C. Kale, N. Shah, K. Jung, Learning effective representations from clinical notes, arXiv preprint arXiv:1705.07025, 2017.
[84] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016) 26094.
[85] N. Limsopatham, N. Collier, Adapting phrase-based machine translation to normalise medical terms in social media messages, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 1675–1680, , https://doi.org/10.18653/v1/D15-1194.
[86] S. Henry, C. Cuffy, B.T. McInnes, Vector representations of multi-word terms for semantic relatedness, J. Biomed. Inform. 77 (2018) 111–119.
[87] M. Hughes, I. Li, S. Kotoulas, T. Suzumura, Medical text classification using convolutional neural networks, Stud. Health Technol. Informatics (2017) 246–250.
[88] Y. Zhang, O. Zhang, Y. Wu, H.-J. Lee, J. Xu, H. Xu, K. Roberts, Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge, J. Biomed. Inform. 75 (2017) 129–137, https://doi.org/10.1016/j.jbi.2017.06.014.
[89] Y. Luo, G. Song, P. Li, Z. Qi, Multi-task medical concept normalization using multi-view convolutional neural network, in: AAAI, 2018.
[90] S. Wang, R. Koopman, Semantic embedding for information retrieval, in: BIR@ECIR, 2017, pp. 122–132.
[91] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, N. Elhadad, Multi-label classification of patient notes: case study on icd code assignment, Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[92] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, J. Wang, Automated icd-9 coding via a deep learning approach, IEEE/ACM Trans. Comput. Biol. Bioinf. (2018) 1, https://doi.org/10.1109/TCBB.2018.2817488.
[93] I. Banerjee, M.C. Chen, M.P. Lungren, D.L. Rubin, Radiology report annotation using intelligent word embeddings: applied to multi-institutional chest ct cohort, J. Biomed. Inform. 77 (2018) 11–20.
[94] S. Moen, T.S.S. Ananiadou, Distributional semantics resources for biomedical text processing, in: Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan, 2013, pp. 39–43.
[95] Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, Biowordvec, improving biomedical word embeddings with subword information and mesh, Sci. Data 6 (2019) 52.
[96] F. Godin, B. Vandersmissen, W. De Neve, R. Van de Walle, Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations, Proceedings of the Workshop on Noisy User-generated Text, 2015, pp. 146–153.
[97] 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).
[98] R. Chalapathy, E.Z. Borzeshi, M. Piccardi, An investigation of recurrent neural architectures for drug name recognition, arXiv preprint arXiv:1609.07585, 2016.
[99] J.L. Elman, Finding structure in time, Cognit. Sci. 14 (1990) 179–211.
[100] M.I. Jordan, Serial order: a parallel distributed processing approach, Advances in psychology, vol. 121, Elsevier, 1997, pp. 471–495.
[101] A.L. Beam, B. Kompa, I. Fried, N.P. Palmer, X. Shi, T. Cai, I.S. Kohane, Clinical concept embeddings learned from massive sources of multimodal medical data, arXiv preprint arXiv:1804.01486, 2018.
[102] A.R. Aronson, F.-M. Lang, An overview of metamap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (2010) 229–236.
[103] S.G. Finlayson, P. LePendu, N.H. Shah, Building the graph of medicine from millions of clinical narratives, Sci. Data 1 (2014) 140032.
[104] S. Wang, R. Koopman, Semantic embedding for information retrieval, 2017.
[105] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, G.B. Melton, Semantic similarity and relatedness between clinical terms: an experimental study, in: AMIA annual symposium proceedings, volume 2010, American Medical Informatics Association, 2010, p. 572.
[106] S.V. Pakhomov, T. Pedersen, B. McInnes, G.B. Melton, A. Ruggieri, C.G. Chute, Towards a framework for developing semantic relatedness reference standards, J. Biomed. Inform. 44 (2011) 251–265.
[107] T. Pedersen, S.V. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, J. Biomed. Inform. 40 (2007) 288–299.
[108] A. Hliaoutakis, Semantic similarity measures in mesh ontology and their application to information retrieval on medline, Master's thesis, 2005.
[109] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, H. Liu, A comparison of word embeddings for the biomedical natural language processing, J. Biomed. Inform. 87 (2018) 12–20, https://doi.org/10.1016/j.jbi.2018.09.008.
[110] D. Newman-Griffis, A. Zirikly, Embedding transfer for low-resource medical named entity recognition: A case study on patient mobility, in: Proceedings of the BioNLP'18 workshop, Association for Computational Linguistics, 2018, pp. 1–11.
[111] B. He, Y. Guan, R. Dai, Classifying medical relations in clinical text via convolutional neural networks, Artif. Intell. Med. (2018), https://doi.org/10.1016/j.artmed.2018.05.001.
[112] F. Qian, C. Gong, L. Liu, L. Sha, M. Zhang, Topic medical concept embedding: Multi-sense representation learning for medical concept, Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 404–409, , https://doi.org/10.1109/BIBM.2017.8217683.
[113] M. Faruqui, J. Dodge, S.K. Jauhar, C. Dyer, E. Hovy, N.A. Smith, Retrofitting word vectors to semantic lexicons, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1606–1615.
[114] Z. Yu, T. Cohen, B. Wallace, E. Bernstam, T. Johnson, Retrofitting word vectors of mesh terms to improve semantic similarity measures, Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, 2016, pp. 43–51.

[115] Z. Yu, B.C. Wallace, T. Johnson, T. Cohen, Retrofitting concept vector representations of medical concepts to improve estimates of semantic similarity and relatedness, Stud. Health Technol. Informatics 245 (2017) 657.
[116] M. Alawad, S.S. Hasan, J.B. Christian, G. Tourassi, Retrofitting word embeddings with the umls metathesaurus for clinical information extraction, 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 2838–2846.
[117] H. Jo, S.J. Choi, Extrofitting: Enriching word representation and its vector space with semantic lexicons, in: Proceedings of The Third Workshop on Representation Learning for NLP, 2018, pp. 24–29.
[118] K. Roberts, Assessing the corpus size vs. similarity trade-off for word embeddings in clinical nlp, Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), The COLING 2016 Organizing Committee, 2016, pp. 54–63.
[119] J.L. Chiang, M.S. Kirkman, L.M. Laffel, A.L. Peters, Type 1 diabetes through the life span: a position statement of the american diabetes association, Diabetes Care 37 (2014) 2034–2054.
[120] B. Murphy, P. Talukdar, T. Mitchell, Learning effective and interpretable semantic models using non-negative sparse embedding, in: Proceedings of COLING 2012, 2012, pp. 1933–1950.
[121] H. Luo, Z. Liu, H. Luan, M. Sun, Online learning of interpretable word embeddings, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1687–1692.
[122] A. Fyshe, P.P. Talukdar, B. Murphy, T.M. Mitchell, Interpretable semantic vectors from a joint model of brain-and text-based meaning, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2014, NIH Public Access, 2014, p. 489.
[123] S. Arora, Y. Li, Y. Liang, T. Ma, A. Risteski, Linear algebraic structure of word senses, with applications to polysemy, Trans. Assoc. Comput. Linguist. 6 (2018) 483–495.
[124] A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, E. Hovy, Spine: Sparse interpretable neural embeddings, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[125] A. Zobnin, Rotations and interpretability of word embeddings: the case of the russian language, International Conference on Analysis of Images, Social Networks and Texts, Springer, 2017, pp. 116–128.
[126] S. Park, J. Bak, A. Oh, Rotated word vector representations and their interpretability, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 401–411.
[127] L.K. Şenel, I. Utlu, V. Yücesoy, A. Koc, T. Cukur, Semantic structure and interpretability of word embeddings, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (2018) 1769–1779.
[128] Z. Chen, Z. He, X. Liu, J. Bian, Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases, BMC Med. Inform. Decision Making 18 (2018) 65.
[129] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, N.A. Smith, Sparse overcomplete word vector representations, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1491–1500.
[130] J. Ba, R. Caruana, Do deep nets really need to be deep? Adv. Neural Inf. Process. Syst. (2014) 2654–2662.
[131] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531, 2015.
[132] A.C. Kozlowski, M. Taddy, J.A. Evans, The geometry of culture: Analyzing meaning through word embeddings, arXiv preprint arXiv:1803.09288, 2018.
[133] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Adv. Neural Inf. Process. Syst. (2016) 4349–4357.
[134] K.M. Hoffman, S. Trawalter, J.R. Axt, M.N. Oliver, Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites, Proc. Nat. Acad. Sci. 113 (2016) 4296–4301.
[135] A. Bakarov, A survey of word embeddings evaluation methods, arXiv preprint arXiv:1801.09536, 2018.
[136] M. Faruqui, Y. Tsvetkov, P. Rastogi, C. Dyer, Problems with evaluation of word embeddings using word similarity tasks, Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, 2016, pp. 30–35.