May 2023

Leveraging Biomedical Ontological Knowledge to Improve Clinical Term Embeddings

Fuad Hatem Abuzahra
University of Wisconsin-Milwaukee

LEVERAGING BIOMEDICAL ONTOLOGICAL KNOWLEDGE TO IMPROVE CLINICAL TERM EMBEDDINGS

by

Fuad Abu Zahra

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

May 2023

ABSTRACT

LEVERAGING BIOMEDICAL ONTOLOGICAL KNOWLEDGE TO IMPROVE CLINICAL TERM EMBEDDINGS

by
Fuad Abu Zahra

The University of Wisconsin-Milwaukee, 2023
Under the Supervision of Dr. Rohit J. Kate

This research is on obtaining and using word embeddings for natural language processing tasks in the biomedical domain. Word embeddings are vector representations of words commonly obtained from large text corpora. This research leverages the biomedical ontology of SNOMED CT as an alternate source for obtaining embeddings for clinical terms. The existing graph-based methods can only give embeddings for concepts (i.e., nodes of the graph) of an ontology, hence we developed a novel method to obtain embeddings for clinical words and terms from their concept embeddings. These embeddings were evaluated on benchmark datasets of clinical term similarity and on the clinical term normalization task and were found to work better than corpus-based embeddings.

However, unlike corpus-based embeddings, the embeddings obtained from SNOMED CT do not incorporate linguistic knowledge as the method was not trained on text data. Therefore, we also developed two new methods to combine the two resources of embeddings – by generating a synthetic corpus out of SNOMED CT ontology and using it for additional training using corpus-based methods, and by fine-tuning a corpus-based system on SNOMED CT concept embeddings. The evaluation showed that the combined embeddings obtained using these methods perform better than either type of embeddings.

To my beloved family,

without whom none of my success would be possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Problem Statement and Motivation

In natural language processing, neural networks are rapidly becoming the standard technology. Given that neural networks require words presented in numerical form, the distributional representation of words, also known as word embedding, is receiving a lot of attention as a result. Word embeddings are the vectorial representations of words such that words with similar meanings and linguistic properties have similar vectors.

The most common source of obtaining word embeddings has been large text corpora. Corpus based methods derive word embeddings using the fact that similar words would appear in similar contexts in text [22][23][31][36]. However, an alternate source for obtaining word embeddings are ontologies which have been relatively less explored. As an example, “pneumonia” and “pneumoconiosis” are both inflammatory disorders of lungs and hence are similar in meaning. This information will be readily available in an ontology of clinical concepts such as SNOMED CT. But to learn this similarity through corpus-based methods, these two terms will need to occur in similar contexts multiple times in a text corpus which may not always happen. In the general natural language processing domain, WordNet [11] is an ontology that arranges words in a graphical structure based on their meanings and relations. Recently, WordNet has been successfully utilized to obtain word embeddings [23]. In the clinical domain, SNOMED CT is an ontology of clinical concepts, and each clinical concept has one or more clinical terms associated with it. Unlike words in WordNet, clinical terms in SNOMED CT are typically multi-word, e.g., "chronic obstructive pulmonary disease" and "acute respiratory distress syndrome." This makes it challenging to obtain word embeddings from the SNOMED CT graph. Another challenge is that the graph methods give embeddings for clinical concepts (nodes in the graph) and not clinical terms/words.

In this study, we developed methods for leveraging the biomedical ontology of SNOMED CT as an alternate source for obtaining embeddings for clinical terms. The current graph-based methods

provide embeddings for the concepts; we developed a novel method to obtain embeddings for clinical words and terms. To the best of our knowledge, this is the first such method. The method can also give embeddings for new clinical terms which are not present in SNOMED CT. The embeddings obtained by this method were evaluated on clinical term similarity and normalization tasks. Although good at capturing ontological knowledge, the SNOMED CT based embeddings do not capture linguistic knowledge, for example, knowledge about how the terms and words are used in sentences. This is because the method was never trained on sentences in a text corpus. Hence we developed two new methods that obtain embeddings by combining the two resources of obtaining embeddings. The first method fine-tunes a corpus-based embedding system by using concept embeddings as targets and thus incorporates ontological information into corpus-based embeddings. The second method generates a synthetic corpus of full sentences out of the SNOMED CT ontology which is when used as additional text corpus to train corpus-based method. The evaluation showed that the combined embeddings obtained using these methods perform better than either type of embeddings.

1.2 Research Objectives and Questions

The research questions that we investigated and answered in our research are:

1. Can the ontological knowledge of clinical concepts in SNOMED CT be utilized to obtain embeddings for clinical concepts and terms? How well will they work, and how will they compare with embeddings obtained using corpus-based methods?
2. What are the best methods for obtaining clinical concept embeddings leveraging all aspects of the SNOMED CT's ontological graph? Moreover, what are the best methods for obtaining and composing clinical term embeddings from clinical concept embeddings? Finally, how will these methods compare?

3. Can clinical term embeddings obtained using SNOMED CT be combined with embeddings obtained using corpus-based methods resulting in better embeddings? What are the best ways to combine the embeddings, and how well will they work?

To address the first question, we developed a method that learns embeddings of clinical terms and words from embeddings for medical concepts obtained by graph-based representation learning. Then, we evaluate and compared these embeddings with embeddings obtained using corpus-based methods. To answer the second question, different methods were used to generate clinical concept embeddings from SNOMED CT's ontological graph. To tackle the third question, we used different methods for combining the embeddings and developed two new methods. Both contextual (BERT) and non-contextual (word2vec) corpus-based embeddings were used to combine with SNOMED CT based embeddings and were evaluated.

1.3 Contributions

This study presents new methods to leverage the biomedical ontology of SNOMED CT as an alternate source for obtaining embeddings for clinical terms. However, the current graph-based methods can only provide embeddings for concepts, a novel method was developed to obtain embeddings for clinical words and terms. We use both classic intrinsic tasks, such as semantic similarity and relatedness, and an extrinsic task to evaluate the resulting word embeddings. We also present two new methods to combine the resources of ontology and text data for obtaining embeddings to incorporate ontological as well as linguistic knowledge into embeddings.

2 Background

2.1 Semantic space

The distributional semantic space model of lexical semantics was proposed in the 1950s [3][4], and it draws on Wittgenstein's theory that the semantics of a word are determined by the context in which it appears [5]. As the other two techniques do, it does not represent the words in a graph but rather as vectors in a high-dimensional space. Similar or related words are represented by vectors near together, whereas dissimilar and unconnected words are represented by vectors that occupy separate regions of the space. For example, some close neighbors of a vector representing a bird may include animal, wings, canary, and sparrow vectors, but the distance between a bird and the kitchen, computer, or scarf would be far.

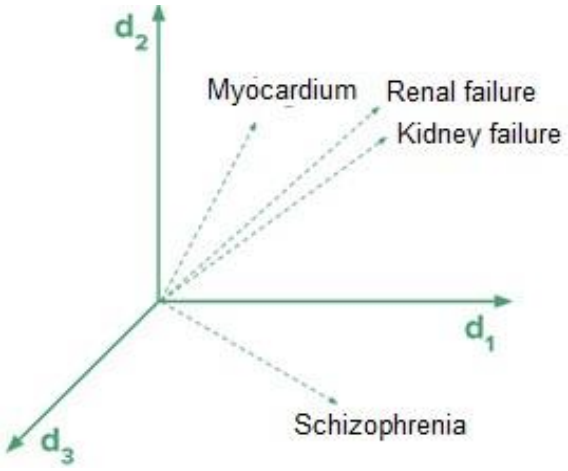


Figure 1 Word embeddings are a type of vectorial representation of words.

An example is the 3-dimensional semantic space schema. The vector representations of the words Renal failure and Kidney failure are highly similar (high level of similarity), with the vector of Myocardium projected further but not too far away (moderate level of similarity) and the vector of Schizophrenia projected in a different area of the space.

2.2 Ontologies

Ontologies are the second type of semantic proxy. These are usually structured information bases that are curated by specialists and are often targeted at a particular domain (e.g., SNOMED CT

[2]). The evidence gathered from the ontology's structure serves as the foundation for object comparison.

Two example ontologies are investigated: an inference-based semantic network (WordNet) and a feature-based network (Small World of Words). Words are grouped into synsets, collections of synonyms that each define a different notion. Conceptual-semantic and lexical linkages are used to link synsets together. Linguists construct such structures; therefore, all existing relationships are curated by experts. These make it an influential and trustworthy source of information for users looking for a comprehensive online thesaurus, computational linguistics, and natural language processing systems.

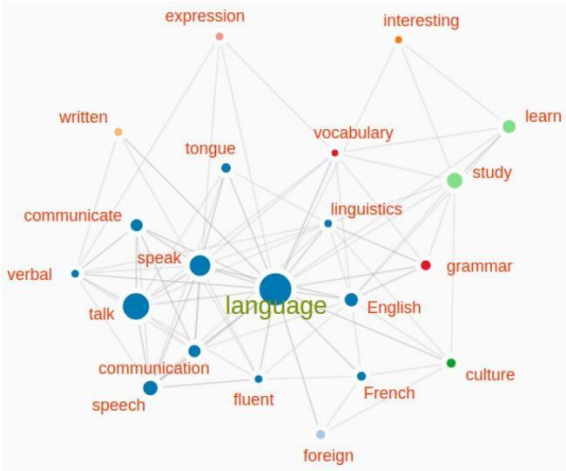


Figure 2 An example entry from WordNet Source:

http://www.smallworldofwords.com/new/visualize

The query word language has produced a subgraph of the Small World of Words. The words given as reactions to the cue language make up the network. The edges represent the association between the words. If there is an edge between two nodes, one of them was given as a response to the other. Source: http://www.smallworldofwords.com/new/visualize

2.3 Word embeddings

The goal of embedding, in general, is to project a collection of objects into a vector space while preserving their essential features. The general idea is to keep the objects' similarity in terms of

distance in the embedding space: comparable objects are closer together while different objects are further apart. As a result, word embedding aims to project words into a semantic space that approximates the distributional semantic space (outlined in Section 2.1) Words (now represented as vectors of numbers) can be processed more efficiently with this representation, especially in neural network-based systems.

2.4 Sources for word embeddings

Multiple research projects have looked into semantic measurements used to compare various aspects of language, including words, sentences, entire documents, and ideas described in knowledge bases [6]. These measurements examine semantic proxies, from which semantic information retrieval will subsequently facilitate object comparison, as the authors of [6] point out. Textual corpora and clinical ontologies are two types of semantic proxies. Experts construct ontologies that include concepts and their relationships. It makes an influential and trustworthy source of information for both users looking for a comprehensive online thesaurus and natural language processing systems.

2.4.1 Textual corpora

The semantic metrics based on textual corpora utilize natural language distributional characteristics, assuming that semantically relevant terms co-occur together. It captures a sense of relatedness between words. Because the words ‘coffee’ and ‘cup’ regularly co-occur in corpora, we can assume they are more semantically connected than, say, ' coffee’ and ‘volcano,' which are unlikely to occur near together. Words found in similar contexts are also deemed similar (e.g., word2vec type methods [22]). The words 'coffee' and 'tea' are semantically related hence will be found in text in similar contexts, i.e., surrounded by similar words.

2.4.2 Clinical ontologies

Ontologies are structured information bases that are curated by specialists and are often targeted at a specific domain (e.g., SNOMED CT [2] in the clinical domain). The evidence gathered from the

ontology's structure serves as the foundation for object comparison. In this study, the ontology investigated is SNOMED CT [2]. Systemized Nomenclature of Medicine—Clinical Terms (SNOMED CT) is a standardized representation of clinical concepts whose extensiveness and expressivity make it suitable for precisely encoding clinical phrases. Next is an example of the SNOMED Ct concept. The following figure shows the design of SNOMED CT.

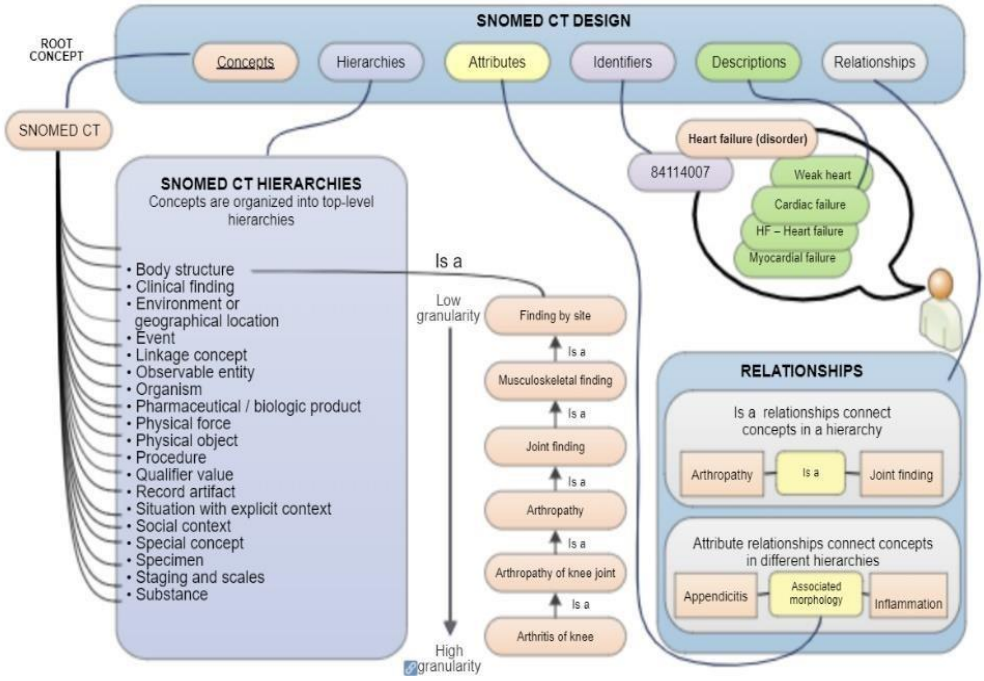


Figure 3 SNOMED CT content and structure [42]

SNOMED CT is a core clinical healthcare terminology that contains concepts with unique meanings and formal logic-based definitions organized into hierarchies. [35] SNOMED CT [2] is the world's most comprehensive clinical oncology, with over a quarter million concepts and more than one and a half million relationships. Its design is a description logic framework [30], allowing automated reasoning. Furthermore, because concepts in SNOMED CT are in their relationships with other concepts, simply recognizing a concept in this ontology can reveal a lot about it, both explicitly stated and implicitly inferred relations.

2.4.3 Relationships and concepts

Every concept in SNOMED CT represents a distinct medical concept and has its own identifier and a few synonym names (AKA descriptions). Although it may not be the clinically chosen term for the notion, one of these is a full-specified name, straightforward design, stable across many contexts, and optimally understood [33]. SNOMED CT has descriptions in various languages, but we exclusively used English descriptions for our study. Each idea also has a semantic type from one of SNOMED CT's nineteen top-level hierarchies, including disorder, finding, and body structure. A concept's semantic type appears inside parenthesis next to its fully described name. The clinical concept of viral meningitis, whose unique identity is 58170007, the fully-specified name is "viral meningitis," semantic type is a disorder, and two alternative descriptions are "abacterial meningitis" and "aseptic meningitis, viral," is depicted in the diagram below.

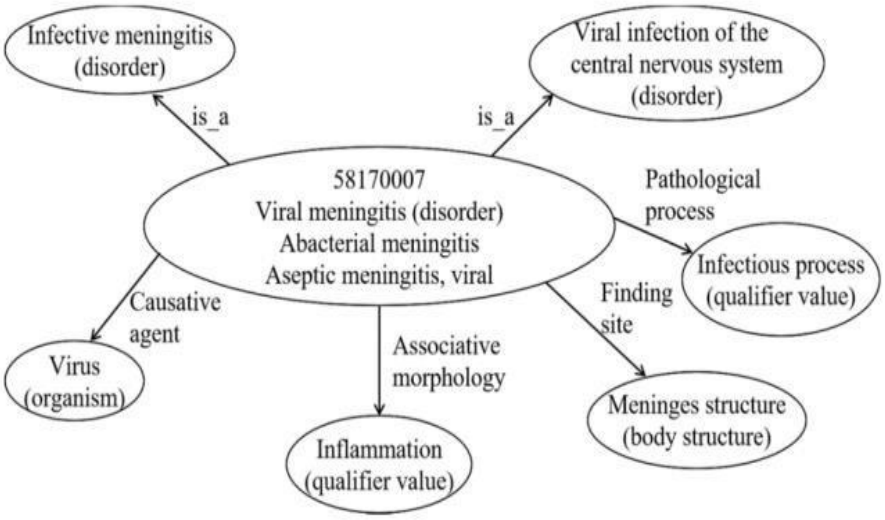


Figure 4 SNOMED CT concepts representation and their relationships [32]

SNOMED CT concepts represent in terms of their relationships with other concepts.

Figure 4 shows a clinical concept called “viral meningitis”, it is an infectious disease, caused by virus, its unique identifier is 58170007, fully-specified name is “viral meningitis”, semantic type is “disorder”, and two other descriptions are “Abacterial meningitis” and “Aseptic meningitis, viral”.

2.5 Contextual embeddings

BERT is a contextual language representation model built on bidirectional transformer encoders [44]. Training a BERT model needs pretraining with vast amount of data to minimize losses for the Masked Language Modelling (MLM) which used to predict the randomly masked tokens and the Next Sentence Prediction (NSP) is used to anticipate the next sentence. Also, the fine-tuning training use moderate number of field specific data and a few epochs for training. The training input sequence starts with a special token [CLS] and uses a [SEP] special token to separate between the sentences in the corpus [43]. The model design involves multiple consecutive layers of an identical architecture; the main component is the multi-head self-attention which computes a contextual hidden representation of each token. This component allows to attend over all positions in the input sequence every position in the decoder, the tokens' embeddings from the last layer can be used as the input of a downstream task [44]. BERT computes different embeddings for the two occurrences of "bank" in the following sentence "the bank robber was seen on the riverbank", unlike the non-contextual model that is biased towards the most frequent meaning in the corpus. [43]

2.6 Related Work

2.6.1 Word embeddings based on textual corpora

Bengio et al. advocated employing neural networks to develop a statistical language model while simultaneously training word embeddings with textual corpora. The authors presented a feedforward neural network comprising an input and projection layer, one hidden and output layer, and one hidden layer. The network was trained and assessed on a language modeling task using a variety of corpora, demonstrating that it outperformed the best available n-gram models. The model, however, was computationally expensive because of the hidden layer's large number of trainable parameters and the SoftMax function's computation.

Because of the word2vec model [22], [23], neural word embeddings have received much traction. The authors proposed two new architectures (CBOW and Skip-gram). The training tasks in the two architectures are different: instead of language modeling (predicting the next word given the n preceding context words), the CBOW model tries to predict the middle word given n context words on the left and right, whereas the Skip-gram model tries to predict the context words given the middle one.

In addition, the authors offered additional model optimizations in [23]. One of them is negative sampling, which replaces the hierarchical SoftMax function (which is an approximation of the full SoftMax). This method saves time and money by avoiding the time-consuming computation of the probability distribution over the vocabulary. Instead, k negative examples are created for each training sample (by randomly selecting words from the vocabulary or using some preset probabilities), and the error is backpropagated just to the weights of those words, not throughout the entire lexicon. Another optimization is frequent word subsampling, which minimizes the quantity of generated training data while reducing the bias towards frequent terms.

These strategies sped up the training process significantly and produced higher-quality embeddings than the model [21]. The cosine similarity of the various vectors is a standard method to compute the similarity of the words in such models, using the formula:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{a}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{b}_i)^2}}$$

a and b are vectors, and ai is the value of vector a's ith coordinate. Thus, the cosine similarity of two vectors oriented in the same direction is 1, and orthogonal vectors are 0, while vectors oriented in opposite directions have a similarity of -1.

Word	Similarity $\in [0, 1]$
languages	0.8418
vocabulary	0.7185
Language	0.6996
spoken	0.6994
grammar	0.6941
linguistic	0.6868
dialect	0.6806
translation	0.6478
English	0.6374
word	0.6259

Table 1 The most similar words to the word ‘language’ [1]

The most similar terms to the word language use the vector similarity in the GloVe embedding model [1]. The previous table shows an example word and the ten most similar words (based on the cosine similarity of the corresponding vectors).

It is worth mentioning that training such models with just a substantial textual sample. This method has several advantages over using ontologies, one of which is that the corpus does not require any form of labeling next section. Furthermore, by feeding the model with additional training corpora, these models can continuously capture the changes in meaning in the language.

It is also worth noting that all terms in the corpora are handled as ambiguous, meaning that no differentiation exists between multiple interpretations of the same word. The resulting vector representations are frequently dominated by a single (most frequent) meaning because the models rely on the statistical properties of word co-occurrences.

2.6.2 Conversion from semantic ontologies to semantic spaces

Related work on obtaining embeddings from ontology is relatively recent. For example, [12] looked into retrofitting to refine distributional representations using relational information, and [13] looked into refining word embeddings using lexical knowledge. However, neither of these papers addressed the goal of obtaining semantic spaces solely based on semantic networks, as we do here.

That is also the goal of recent work like [14], which uses the local, one edge relations of each relevant term in the WordNet network to improve embeddings created from data sets of selected Wikipedia articles. Other recent research worth mentioning includes [15], which used order embeddings to maintain distance and/or directionality under the relevant semantic relations, but not distance and/or directionality under the relevant semantic relations. [16] that used the Poincaré ball model to compute embeddings in hyperbolic space rather than Euclidean space.

In contrast, [17] provides an example of the stability of wnet2vec when plugged into neural models and its implementation in a downstream task, where these embeddings facilitate neural network-based brain activation prediction. There has also been a long legacy of study on learning vector embeddings from multi-relational data, with references to [18][19][20]. Though these are, to a significant degree, generic ways for a graph to vector conversion, the focus here has been on testing these models' capacity to complete missing relations in knowledge bases rather than on natural language processing and lexical semantics.

De Deyne et al. and Goikoetxea et al. are two other comparable approaches worth mentioning. While both use the identical iterative conversion technique, the first focuses on converting a piece of the lexicon represented using a feature-based approach into a semantic space rather than a semantic network. The second resorts to a lossy intermediate "textual" representation: it generates sequences of words by concatenating words visited by random walks over the WordNet; this "artificial text" is a partial and contingent reflection of the semantic network and is used to obtain distributional vectors by resorting to traditional text-based word embedding techniques.

They used skip-gram to train word and idea embeddings, then fine-tuned with a transformer-based BERT architecture in the two-sentence input mode with a classification aim that captures MeSH pair co-occurrence in [7]. Finally, they employed concept correlations to improve static biomedical word embeddings utilizing a transformer architecture that was previously used to generate dynamic embeddings.

Using graph-based representation learning methods on SNOMED-CT, Agarwal et al. [8] proposed learning embeddings for medical ideas. It resulted in a 5-6x increase in "idea similarity" and a 6-20% increase in patient diagnosis. The researchers used the Node2vec, Metapath2vec, and Poincare algorithms to create embeddings. For the patient state prediction models and capturing the node types, Poincaré-based embeddings learned from hierarchical relations were found to work well.

The paper shows 1) how to build the initial large corpus of texts to train the word2vec models, 2) how to use this vector space model to create final SNOMED2Vec vector space model, and 3) how to use the cosine similarity distance to find the most similar concepts, grouping by SNOMED-CT hierarchies. Furthermore, like a vector space model, they employ word embedding to express the descriptive words of the SNOMED Concepts (SNOMED2Vec).

Finally, they propose a collection of concepts to the human specialist to build a tool for codifying clinical reports.

Schultz et al. [9] artificially generated a few large-scale medical term similarity datasets, and show that an annotation analysis with doctors confirms their high quality. Existing datasets for medical term similarity have been proven to be too small to discover significant performance differences between embeddings and similarity metrics used for embedding. On the other hand, significant disparities are shown utilizing their new large-scale datasets. Furthermore, the new datasets reveal how difficult it is for current embeddings to forecast the similarity of non-obvious term pairs, such as semantically similar but lexically distinct phrases and vice versa.

In this study, the obtained embeddings for clinical terms directly from SNOMED CT ontology outperformed on corpus-based methods using the five benchmark datasets mentioned earlier.

The node2vec algorithmic framework for learning continuous feature representations for nodes in networks is proposed [10]. It shows that node2vec outperforms existing state-of-the-art algorithms on multi-label categorization and link prediction in various real-world networks. Furthermore, they discovered that breadth-first search (BFS) could only explore a limited number of neighborhoods.

As a result, breadth-first search BFS is well suited to describing structural equivalences in networks that rely on nodes' immediate local structure. Depth-first search (DFS), on the other hand, can freely explore network neighborhoods, which helps identify homophilous communities at the expense of significant variation.

Saedi et al. [11] offer a methodology for converting semantic networks into semantic spaces using WordNet, and the performance of the resulting embeddings in a prevalent semantic similarity task outperformed word embeddings obtained using corpus-based word2vec method using vast collections of texts.

2.6.3 Word embeddings based on lexical ontologies.

Lexical ontologies are graph representations that nodes represent lexical units (e.g., words or synsets in WordNet) and edges represent semantic relationship between them. As a result, extracting network node embeddings from ontologies is the only way to get word embeddings. [24] gave a detailed assessment of graph embedding methods in a recent paper. The authors present a method taxonomy (based on problem setting, i.e., the algorithm's type of input and output) and an outline of five significant graph categories embedding techniques. This research focuses on three of the most popular node-embedding algorithms: 1) matrix factorization, 2) random walk, and 3) edge reconstruction. These approaches represent the graph in various ways, impacting how attributes are retained in the embedded space. In the following sections, we will go over these three approaches.

2.6.4 Matrix factorization-based methods

These methods use a matrix to represent the graph attributes, which are then factorized to provide node embeddings. The critical distinction is how the input matrix is built (e.g., adjacency matrix, node proximity matrix) and the objective function to be optimized. The Katz index approach represents matrix factorization-based methods ([25], Eq. 7.63).

The idea behind this metric is that the greater the number of pathways connecting two nodes, the closer they are. As a result, we want to count all paths between two nodes. Raise the adjacency matrix m to the power of p to get a matrix with each cell m_{ij} representing the number of pathways of length p between nodes i and j. As a result, we can iteratively acquire these counts:

$$M_G^n = I + \alpha M + \alpha^2 M^2 + \cdots + \alpha^n M^n$$

Where i is an identity matrix and α is a decay factor (between 0 and 1), allowing for down- weighing the influence of long paths. Interestingly, if we extend this formula to an infinite sum, following [17], we can rewrite it in the following way:

$$M_G = \sum_{p=0}^{\infty} (\alpha M)^p = (I - \alpha M)^{-1}$$

This method enables the simulation of paths of any length on the network using only the adjacency matrix but at the cost of a matrix inversion, which is a computationally expensive operation, especially for larger graphs.

We chose a method from this subgroup as a sample of the matrix factorization models since it was successfully applied to WordNet, where the authors demonstrated that the generated embeddings beat the mainstream text-based embeddings in the semantic similarity challenge [26].

2.6.5 Random-walk-based methods.

These methods describe the graph as a set of random walk paths sampled from the graph, which are then used to extract node embeddings using a deep learning algorithm—for example, training a Skip-Gram model across a synthetic corpus or employing recurrent neural networks, such as those based on Long-Short-Term Memory (LSTM) units.

Perozzi el al. presented Deep Walk; a Skip-Gram based approach of embedding nodes in a graph that was used to learn latent representations in social networks. It was further generalized by [27],

who used biasing the random walk to achieve a more flexible notion of the neighborhood between the nodes.

Goikoetxea et al. used a similar strategy to extract word embeddings from WordNet that outperformed or performed comparably to text-based ones on the semantic similarity challenge. The authors also show that combining text- and graph-based embeddings improves the results, implying that the two models contain different semantic information in the embeddings.

2.6.6 Clinical Term Normalization

Sometimes the clinical notes contain nonstandard clinical terms (terms exist in the ontologies). Some physicians use alternative clinical terms, synonyms, or acronyms in their clinical notes. Which makes it necessary to map the nonstandard terms to their standard form while doing semantic analysis [40]. For example, the medical term "diffuse inflammatory reaction" may relate to the Unified Medical Language System (UMLS) ‘diffuse inflammation’ clinical concept or the “inflammation diffuse” clinical concept. Also, the term “allergy to ferrous sulphate," is not a found in the UMLS terms and their synonyms, but the closest term exists in the UMLS is “allergy to ferrous sulfate" [40].

Kate [55,56] utilizes the MCN corpus to build the clinical term normalization system which transforms the clinical terms into their normalized forms using the edit patterns. The former method uses the UMLS synonyms to learn common variations using Levenshtein edit distance and sequence of edits between any two terms. Different normalization components are used, like exact matching, learned edit patterns, sub-concept matching and disambiguation for multiple concepts. The results shown in section 4.7 use only the exact matching in the clinical term normalization in the clinical embedding evaluation.

2.6.7 Contextual embeddings

Over the last few years, ELMo [57] and BERT [58] have presented strong solutions that can provide contextualized word representations. BioBERT [59] trains a BERT model over a

corpus of biomedical research articles sourced from PubMed article abstracts and PubMed Central article full texts[68]. On clinical text, [60] uses a general domain pretrained ELMo model towards the task of clinical text de-identification. [61], released in late February 2019, train a clinical note corpus BERT language model and uses complex task-specific models to yield improvements over both traditional embeddings and ELMo embeddings.

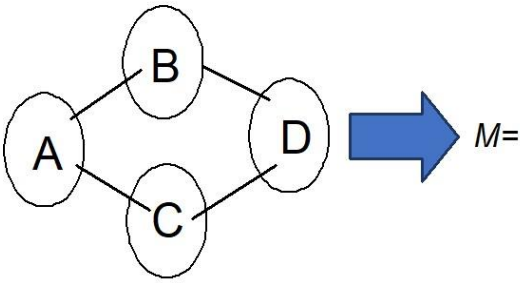
3 Methods

3.1 SNOMED CT matrix factorization embeddings generator

As mentioned in the introduction, this research aims to create clinical term embeddings from SNOMED CT ontology of clinical concepts. Our completed work consists of two parts. The first part is generating embeddings of clinical concepts and clinical terms from SNOMED CT and the second part is to evaluate those embeddings.

The embedding generator program generates embeddings by reading all the SNOMED CT active concepts and relationships into an adjacency matrix of size 365,000 X 365,000, since there are 365,000 concepts in SNOMED CT ontology.

The next figure shows the adjacency matrix for the graph of concepts A, B, C and D. When two concepts are related in the left graph, it will be represented by 1 otherwise 0 in the matrix M in the right side.



M=

	A	B	C	D
A	0	1	1	0
B	1	0	0	1
C	1	0	0	1
D	0	1	1	0

Relation between B and D

Figure 5 Example of adjacency matrix of a graph

Then M2 represents paths of length 2 and so on. $MG = I + \alpha M + \alpha 2\ M2 + \alpha 3\ M3 + ..$ counts all paths weighing down their lengths by a decay factor α (< 1). We found that up to the fifth power of M was sufficient for our purpose.

After that, we reduce the dimensionality of the matrix to different sizes (e.g., 200, 300, 850) using truncated singular value decomposition (SVD). The transformation of data from a high dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data. Each row becomes the concept embedding with 200, 300, or 850 dimensions, making it easier to handle and train in the downstream applications.

3.2 SNOMED CT random walk embeddings generator

The second method is based on the premise of similar terms will have similar "contexts" in the graph, it has been previously used on WordNet [36] and it has been previously applied to SNOMED CT [8] but they did not obtain clinical term embeddings or evaluate them for clinical term similarity. Figure 6 shows an artificial "corpus" by simulating random walks on the graph.

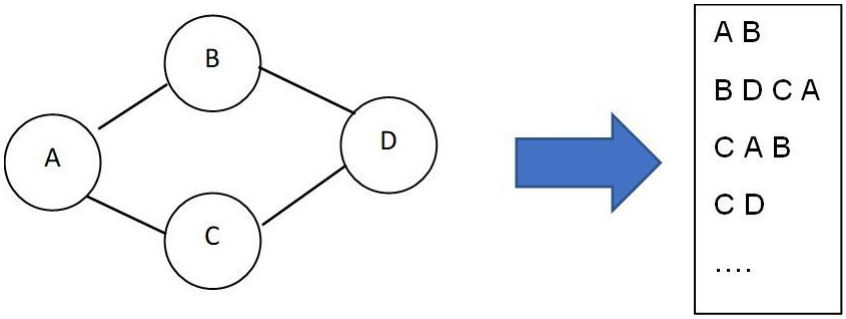


Figure 6 Example of a random walk of a graph

After the artificial corpus of random walk "sentences" has been created, corpus-based methods are used to obtain embeddings of the concepts in the graph. [6].

In this chapter, we will present this method for generating relationships embeddings for SCOMED CT medical ontology. Figure [7] shows the overall process to our proposed method using March 2022 SNOMED CT files which include concepts file "sct2_Concept_Full_ US1000124_20220301 .txt" and the relations file "sct2_Relationship_Full_US1000124_20220301 .txt" and the descriptions file "sct2_Description_Full-en_US1000124_20220301.txt".

In this approach, we build four main dictionaries: 1- active_concepts, 2- active_descriptions, 3- adjacents_concepts, 4- random_walks.

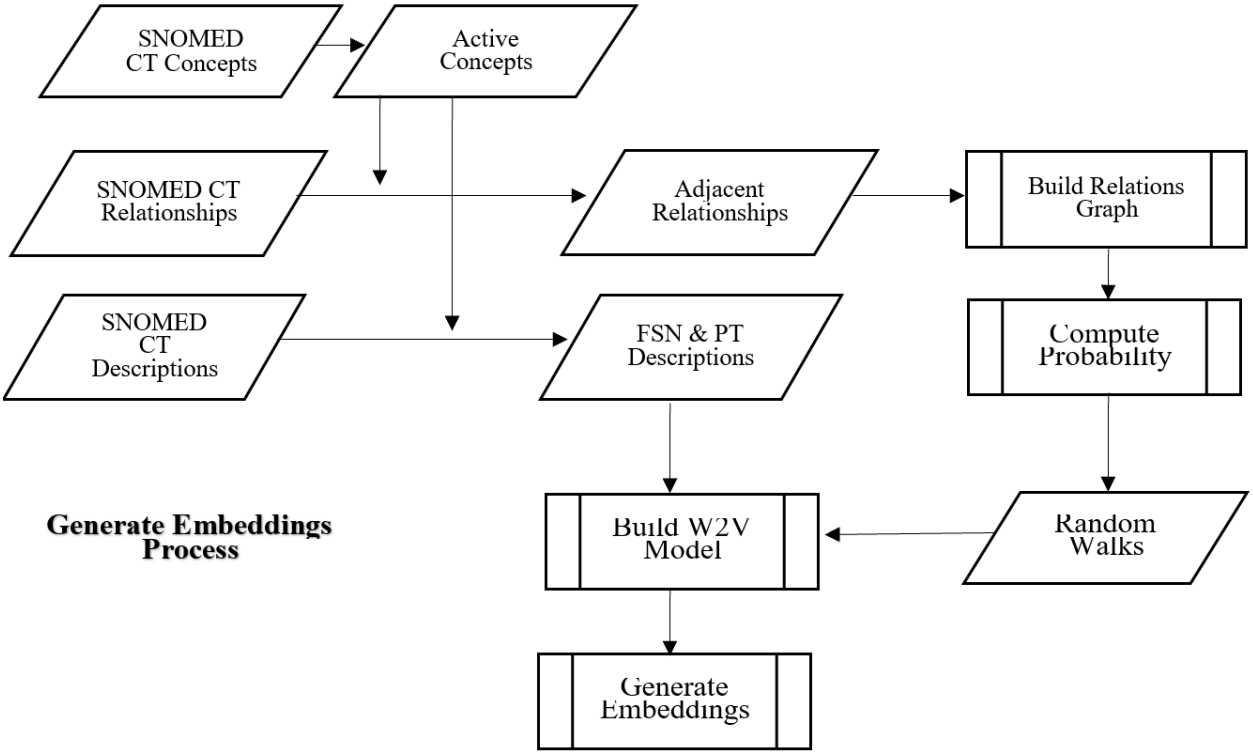


Figure 7 SNOMED CT Embeddings Generator using Random Walk

3.3 Clinical Term Embeddings Generator

The methods described in the previous sections can obtain only clinical concept embeddings, however, for NLP applications one needs embeddings for words as well as clinical terms. In this section, we present a new method for obtaining embeddings for both words and new terms from concept embeddings. We train a deep neural network to predict the concept embeddings from clinical terms. It is a recurrent neural network with two layers of GRU [45]. The neural network diagram in figure 8 below uses concept embeddings as learned by matrix factorization method or random walk method as targets; and uses the corresponding clinical terms listed in SNOMED CT as well as their UMLS synonyms as inputs. The embedding layer learns the embedding for words and can be read off from a trained network. The trained network can then also predict embeddings for new clinical terms as shown in figure 9.

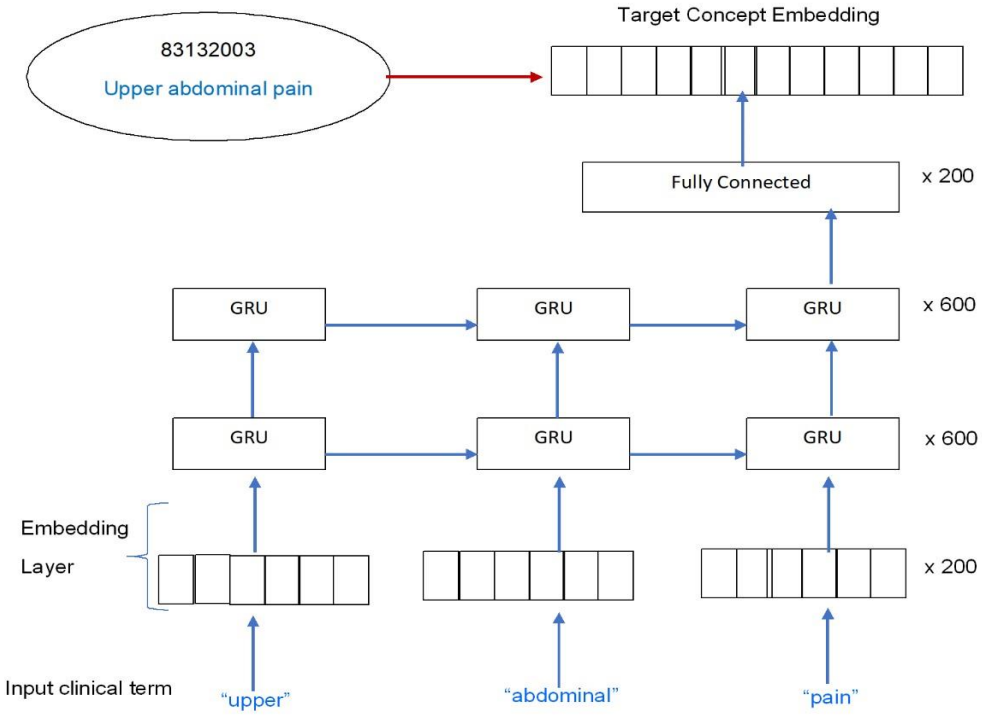


Figure 8 The Clinical Term Embeddings Generator. A deep Neural Network use Random Walk embeddings as an input for the SNOMED CT concept's Descriptions

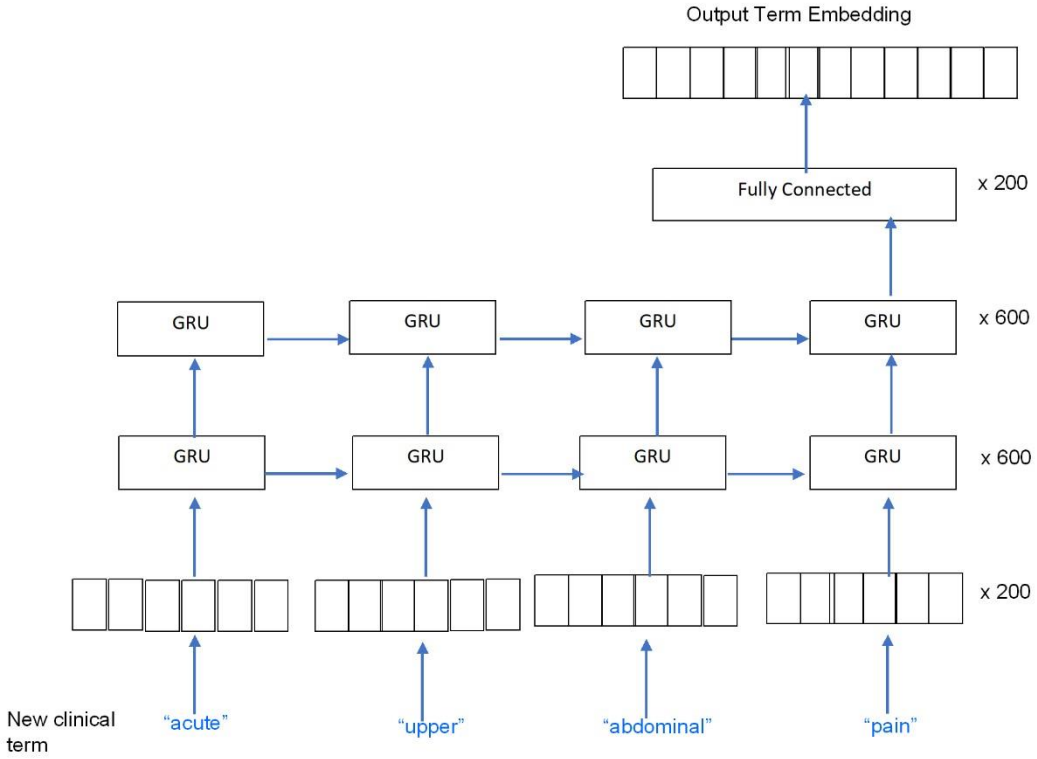


Figure 9 Predicting a new medical term using the Clinical Term Embeddings Generator

3.4 Clinical term normalization

In this task we evaluated clinical term embeddings obtained from SNOMED CT is the clinical term normalization task [53]. In this task, given a clinical term, it is to be mapped to its concept (identified by an identifier) in a medical terminology, typically in UMLS Metathesaurus [54]. As shown in figure 10:

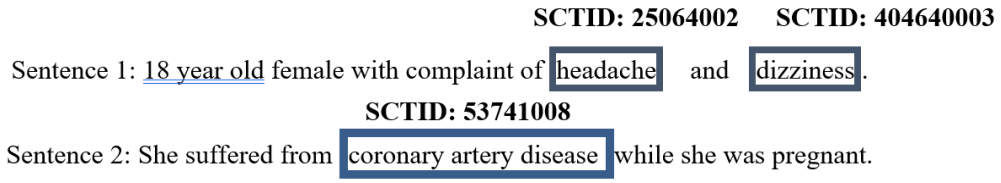


Figure 10: Clinical Term mapped to its concepts in a medical terminology.

For our experiments, we used the benchmark MCN dataset [34] which has been used extensively for evaluating normalization methods [53]. This dataset has 6,684 clinical terms for training and 6,925 clinical terms for testing. In the entire dataset, 2.7% of clinical terms are "CUI-less", that is, they do not correspond to any concept in UMLS, while others are paired with their correct concept unique identifiers (CUIs).

To normalize a clinical term, our method first tries to exactly match it in UMLS as well as in the training examples. If it exactly matches, then the concept corresponding to that term is given as the output. In case it matches multiple clinical terms corresponding to multiple concepts then a method described later is used to disambiguate the clinical term. If the clinical term does not match exactly either in the training data or in UMLS, then the method first obtains embedding of the clinical term using the model described earlier. It then computes cosine similarity of this embedding with the embedding of every clinical term in UMLS and determines the closest clinical term. The concept corresponding to this closest clinical term is then given as the output. For efficiency, the embeddings of all the clinical terms in UMLS are pre-computed using the model. Through pilot experiments we found that besides cosine similarity, including the fraction of the words common between the two clinical terms is also useful, especially when the terms have rare

words in common for which good embeddings may not have been learned by the model. We define similarity between two clinical terms as weighted similarity with 90% weight of the cosine similarity and 10% weight of the fraction of the words common between them.

We observed a limitation of the embeddings obtained from SNOMED CT which is also a limitation of corpus-based methods. The model learns very similar embeddings for words with opposite meanings, for example, “left” and “right”, or “acute” and “chronic”. In addition, it learns similar embeddings for words with analogous meanings but that completely change the meaning of a clinical term, for example, “primary” and “secondary”, or “cervical” and “thoracic”. This happens because the clinical terms with opposite or analogous meanings will have their concepts in very similar positions in the ontological graph. As a result, our model tends to learn very similar embeddings for such words even though they change the meanings of clinical terms. Corpus-based embeddings also suffer from this limitation because words with opposite or analogous meanings are often found in similar contexts in text.

Another limitation we observed was that the model sometimes would learn different embeddings for words with similar meanings which could be synonyms or sometimes spelled differently, for example, “ultrasonography” and “ultrasound”, or “edema” and “oedema”, or “bilateral” and “left and right”. Because the method considers each word separately, it may not learn the same embeddings for them. This affects normalization when the given term is, for example, “left and right kidneys” which then may not normalize to “bilateral kidneys”. This limitation also affects corpus-based embeddings unless they see these words in similar contexts frequently enough.

To counter the above limitations, we augmented our normalization method with some patterns which were automatically learned from UMLS. In the next section, we include results of an ablation study that shows how much they contributed to the normalization task. A pattern is derived from two clinical terms and consists of two parts. The first part consists of words which are present in the first clinical term but not in the second clinical term, and the second part consists of the vice-

versa. For example, given the terms “primary neoplasm” and “secondary neoplasm”, the pattern derived from them will be “primary-secondary”. The two parts of the pattern (shown separate by “-”) are considered inter-changeable when being applied. One of the parts could also be empty which would capture whether presence of an extra word changes the meaning or not. For example, presence of “nos” (not specified) does not change the meaning, but presence of “infected” changes the meaning.

Our pattern learning method efficiently considers every two clinical terms in UMLS and determines all the patterns and their number of positive and negative matches. If a pattern matches two clinical terms which share the same concept, then it is considered a positive match, otherwise it is considered a negative match implying that the clinical terms mean different things. Ambiguous clinical terms (that are associated with more than one concept) are not included in this learning process. To make the process efficient, only those pair of clinical terms are considered which have at least half the words in common. To avoid large patterns that may not match often, the patterns were restricted to have the combined length of the two parts to be less than five. We call the patterns negative patterns if they have more than 5 negatives and 10 times more negatives than positives. As an example, “anterior-posterior” is a negative pattern. Similarly, we call the patterns positive patterns if they have more than 5 positives and 10 times more positives than negatives. As an example, “bilateral-left and right” is a positive pattern. These patterns are different from patterns from past work [55,40], because those patterns were meant to generate a new clinical term with the same meaning and could not handle clinical terms with opposite or analogous meanings. In contrast, these patterns are meant to determine if two clinical terms represent the same concept or different concepts.

These learned patterns are used in the normalization method as follows. In the first step of normalization, if the clinical term exactly matches multiple clinical terms in UMLS corresponding to multiple concepts then the patterns are used to disambiguate from these candidate concepts. If

a negative pattern matches the pair of the given clinical term and any clinical term corresponding to one of the candidate concepts, then that concept is removed as a candidate. On the other hand, if a positive pattern matches then that concept is given as the output (unless the same happens with another candidate concept). If after applying the patterns, more than one concepts remain then the average similarity between the given clinical term and all the clinical terms in UMLS corresponding to that concept is computed and the concept with the highest similarity is given as the output.

If the clinical term does not exactly match in UMLS, then all clinical terms in UMLS are considered. If a positive pattern matches a clinical term (paired with the given clinical term) then the concept corresponding to it is given as the output (unless there are more than one such concept) with the similarity score of 1. But if a negative pattern matches then that concept can never be the output. If no positive pattern matches, then similarity is computed with all the clinical terms. The most similar term is given as the output. However, if the difference between the similarity of the top concepts is too close (less than 0.001) then the average similarity with all the clinical terms in UMLS corresponding to the top concepts are considered and used to determine the most similar concept. If no clinical term is found in UMLS with more than 0.9 similarity, then "CUI-less" is given as the output.

3.5 Combined Embeddings

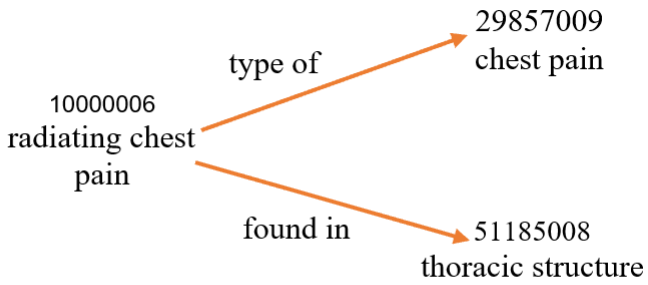
3.5.1 Concatenate Embeddings

The first method we used for combining SNOMED CT based embeddings with corpus-based embeddings was by simply concatenating them. In this method, we tried to improve our embeddings capabilities by adding extra information from general purpose sources like GoogleNews medical terms or the medical words used in Wiki documents. Tomas Mikolov created it at Google in 2013 to make neural network-based embedding training more efficient; ever since it seems to be everyone's favorite pre-trained word embedding. The Google News

dataset was used to train Word2Vec (about 100 billion words!). This repository hosts the word2vec pre-trained Google News corpus (3 billion running words) word vector model (3 million 300-dimension English word vectors). In [10], Wang et. al. used Google new as benchmark dataset for their method which was depending mainly on the general-purpose embeddings.

3.5.2 SNOMED CT corpus generator

We created a new medical corpus from SNOMED CT ontology based on the concept descriptions and relation types of Figures 11. We used the context free grammars CFG from Natural Language Toolkit NLTK library which is often used to find possible syntactic structures for sentences. The CFG class is used to encode context free grammar. Each CFG consists of a start symbol and a set of productions. The "start symbol" specifies the root node value for parse trees. We used part of the relation types, we created more than 1.6M sentences which are used for training a pretrained BERT model and corpus-based model (word2vec).



radiating chest pain is type of chest pain.
radiating chest pain is found in thoracic structure.

Figure 11 SNOMED corpus Generator

Context Free Grammars	Example
S -> VP	The entire stylomastoid foramen is type of structure of stylomastoid foramen.
NP -> Det src_concept_Det N	
VP -> Det src_concept V N	
N -> des_concept	
Det -> the	
V -> rel_txt	
PP -> P -> NP	
P -> 'or'	

Table 2 Context Free Grammars (CFG)

SNOMED CT generated corpus sample	Grammar verb
The parapsoriasis guttata is type of chronic skin disease.	"is type of"
The parapsoriasis guttata is type of chronic dermatosis.	"is type of"
The postnatal examination finding occurs during the maternal pregnancy period.	"occurs during the"
The pulsed electromagnetic energy to back is type of procedure on back.	"is type of"
The procedure on muscle of hand is required for the muscle of hand.	"is required for"
The chronic lichenoid pityriasis is found in skin.	"is found in"
Parapsoriasis guttata is found in skin.	"is found in"
The parapsoriasis guttata is found in skin structure.	"is found in"
The entire suprahyoid region is type of suprahyoid region, the entire suprahyoid region is type of structure of suprahyoid region.	"is type of"
The aspiration of pancreas is required for the pancreatic structure.	"is required for"

Table 3 SNOMED CT generated corpus Sample

3.5.3 Merging Clinical corpus and SNOMED CT generated corpus

We used the SNOMED generated corpus to train a clinical BERT model, the generated embeddings are a combination of large medical corpus from PubMed and PMC, in addition to the SNOMED CT generated corpus. We selected the best BERT hyperparameters (train

steps=100,000, batch size=64, maximum sequence length=128, maximum predictions per sequence=20, learning rate=2e-5).

3.5.4 Fine Tuning a Pretrained BERT

BERT Base (uncased) is a pretrained model on English language using a masked language modeling (MLM) objective [58]. The BERT model was pretrained on a dataset consisting of 11,038 unpublished books and English Wikipedia.

Fine-tuning BERT involves taking the pre-trained BERT model and training it further on a specific downstream task with task-specific labeled data. The idea is to leverage the general language understanding abilities learned by BERT during pre-training and adapt them to the specific task at hand. We used "bert_base_uncased" clinical model. It was then fine-tuned to predict SNOMED CT concept embeddings obtained using graph-based method. SNOMED CT descriptions were used as inputs and the concept embeddings as targets as shown in Figure 13. This process makes BERT incorporate ontological knowledge into its clinical term embeddings. Table 4 shows the hyper parameters used in fine tuning process.

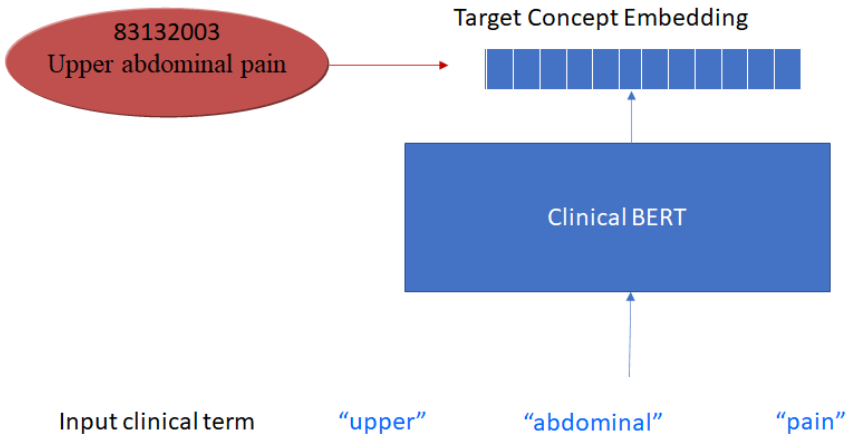


Figure 13: Fine-tuning ClinicalBERT with SNOMED CT concept embeddings to combine corpus-based and ontology-based embeddings.

Batch_Size	32
Learning Rate	1e-3
Num_Epochs	10
Num_Threads	1
Max_Len_Train	128
Max_Len_Valid	128

Table 4 Hyperparameters of the BERT-base-uncased fine tuning. Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning.

4 Results

In this chapter, all the experimental evaluation results are presented.

We evaluate the performance of embeddings obtained using SNOMED CT using different methods on the clinical term similarity tasks. We then present results on the clinical term normalization task. This is followed by the results obtained by combining the corpus-based and SNOMED CT embeddings obtained by different methods.

4.1 Clinical Term Similarity Data Set

In this study, we have used a few data sets from several sources:

The first data set used in this research is Pedersen's dataset [39]. Pedersen dataset is a set of thirty term pairs annotated by three rheumatology physicians. The annotation of each pair is out of a 4-point scale, the average correlation between physicians is 0.68, and the coders is 0.78. In Table 5 “Renal failure” and “Kidney failure” similarity is 4 which means they are highly related to each other, where “Appendicitis” and “Osteoporosis” similarity is 1 that means they are dissimilar.

Term1	Term2	Similarity
Renal failure	Kidney failure	4
Heart	Myocardium	3.3
Acne	Syringe	2
Appendicitis	Osteoporosis	1

Table 5 Sample of Pedersen's data. A set of 30 concept pairs that were then annotated by three physicians and a subset of 9 medical coders. Each pair was annotated on a 4-point scale: practically synonymous (4.0), related (3.0), marginally related (2.0) and unrelated (1.0).

UMNSRS [26] is the second dataset, it consists of 566 medical term pairs, and it is compiled by first selecting all concepts from the UMLS with one of three semantic types: disorders, symptoms, and drugs. Subsequently, only concepts with entry terms containing at least one single-word term were further selected to control for potential differences in similarity and relatedness responses due to differences in term complexity. After this automatic selection, a practicing physician manually

selected pairs of the single-word terms to contain approximately thirty term pairs in each of the four relatedness categories and six semantic type categories of term pairs (DISORDER-DISORDER, DISORDER SYMPTOM, DISORDER-DRUG, SYMPTOM SYMPTOM, SYMPTOM-DRUG, DRUG-DRUG). With terms denoting medications, we used brand names in most cases because generic names for drugs with similar chemical composition and/or function tend to have similar orthography and pronunciation, presenting a potential source of bias.

Term1	Term2	Similarity
Glaucoma	Fibrillation	100.081
Carbatrol	Dilantin	139.8451
Cardiomyopathy	Tylenol	148.9565
Herpes	Hyperthyroidism	94.05415

Table 6 Sample of UMNSRS data. A set of 566 UMLS concept pairs manually rated for semantic similarity using a continuous response scale.

The third dataset MAYOSRS [38], consists of 101 medical term pairs. The ratings are not uniformly distributed for either of the datasets. For the medical term pairs, a more significant proportion of ratings on average are found in the "related" (lower values) than the "unrelated" end of the scale. The distribution for the general English word pairs is bimodal, suggesting that the raters tended to make binary decisions.

Term1	Term2	Similarity
difficulty walking	antalgic gait	6.69
rheumatoid nodule	lung nodule	2.38
hand splint	splinter hemorrhage	1
diabetes	polyp	1

Table 7 Sample of MAYOSRS data. A set of 101 medical concept pairs manually rater by medical coders for semantic relatedness.

The fourth dataset is the Hliaoutakis dataset [37], consisting of 34 medical term pairs with similarity scores obtained by human judgments.

Term1	Term2	Similarity
Anemia	Appendicitis	0.031
Dementia	Atopic Dermatitis	0.062
Bacterial Pneumonia	Malaria	0.156
Osteoporosis	Patent Ductus Arteriosus	0.156

Table 8 Sample of Hliaoutakis data. A set of 49 pairs. Their similarity was evaluated by doctors, giving a score to each pair between 0 (not similar) and 4 (perfect similarity). The average rating (by all doctors) of each pair represents an estimate of how similar each pair is according to human judgement.

The fifth dataset is EHR-RelB, it is an open source of novel concept relatedness benchmark; it is six times larger than existing datasets and the concept pairs are chosen based on the cooccurrence in EHR system [41]. There are 3630 concept pairs sampled from electronic health records (EHRs) sorted uniquely in descending order.

Term 1	Term2	Similarity
Chronic obstructive lung disease	Chronic cor pulmonale	2
Cramp in lower leg associated with rest	Peripheral vascular disease	1.67
Varicose veins of lower extremity	Varicose vein operation	2.67
Retained placenta	Postpartum haemorrhage	2
H/O: regular medication	Allergy to cat dander	0.33

Table 9 Sample of EHR-RelB data. A biomedical concept relatedness dataset consisting of 3630 concept pairs. Dataset is sampled from EHRs to ensure concepts are relevant for the EHR concept retrieval task.

4.2 Medical questions pairs Data Set

Medical questions pairs dataset [69] consists of 3048 similar and dissimilar medical question pairs hand-generated and labeled by Curai's doctors. Doctors with a list of 1524 patient-asked questions randomly sampled from the publicly available. Each question results in one similar and one different pair. Table [8]

Dr_Id	Question 1	Question 2	Label
1	Am I overweight (192.9) for my age (39)?	What diet is good for losing weight? Keto or vegan?	0
1	Does age increase the severity of eds symptoms/problems?	I think I have ED. Would this worsen as I grow old?	1
1	How do I know if I have depression and anxiety?	What are the symptoms of depression and anxiety?	1
1	Can doxycycline treat an ear infection?	What are the side effects of Doxycycline?	0

Table 10 Sample medical questions pairs. [69] A question pairs relatedness dataset consisting of 3048 question pairs. The dataset contains dr_id, question_1, question_2, label. 11 different doctors were used for this task so dr_id ranges from 1 to 11. The label is 1 if the question pair is similar and 0 otherwise.

4.3 Evaluation Measures

The cosine similarity of the various vectors is a standard method to compute the similarity of the words in such models, using the formula:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{a}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{b}_i)^2}}$$

a and b are vectors, and ai is the value of vector a's i^{th} coordinate. Thus, the cosine similarity of two vectors oriented in the same direction is 1, and orthogonal vectors are 0, while vectors oriented in opposite directions have a similarity of -1. For a list of clinical term pairs, the similarities thus computed are compared against the expert-judged similarity scores in the dataset using a measure of correlation coefficient. The computed numbers of Spearman's ranked correlation coefficient between the similarity scores obtained from embeddings and the expert-judged similarity scores. The other computed numbers are Pearson correlation coefficient between the similarity scores obtained using the embeddings and the expert-judged similarity scores.

4.4 SNOMED CT Embeddings Evaluated on Clinical Term Similarity Task

The method to generate clinical term embeddings described in previous sections is evaluated on benchmark datasets of medical terms pair similarity.

4.4.1 Matrix Factorization Method

We use SNOMED CT files to extract the concepts and relations between these concepts, to build the n x n co-occurrence matrix where n is the number of concepts read from the SNOMED CT file. The selected matrix factorization method simulates an infinite random walk by computing the Katz index on the adjacency matrix. This operation involves the inversion of the matrix, which is computationally very expensive and thus, can be challenging for larger graphs. Then we factorized the adjacency matrix to obtain node embeddings, and because the larger the number of paths that exist between two nodes, the more similar they are, we count all the existing paths between two given nodes using the general formula: $MG = (1-\alpha M)^{-1}$.

With vector dimension d=850, the dimensionality reduction using PCA, the vector dimensions are sorted by descending variance and the first n-dimensions are retained as embeddings. There are many algorithms for dimensionality reduction, PCA, IncrementalPCA, TruncatedSVD, and SparsePCA. We used IncrementalPCA for large number of concepts. Finally, we created the embedding file which consists from the concept ID and the 850 embeddings.

We used the generated embedding file in RNN model, then we found similarity correlations for four datasets: Pedersen, UMNSRS, MYOSRS and Hliaoutakis.

For intrinsic evaluation, we evaluate the embeddings in semantic similarity and relatedness tasks, where the similarity of the vectors is matched against gold standard scores established by humans. We got some descent results in Table 9.

	Matrix Factorization Embeddings	
	Word2Vec model	RNN model
Pederson	0.454	0.603
Mayosrs	0.283	0.294
Umnsrs	0.423	0.423
Hliaoutakis	0.622	0.629
EHR-RelB	0.24	0.410

Table 11 Matrix Factorization Embeddings. Pearson correlation coefficient between similarity scores from human judgments and those from word embeddings on four measurement datasets.

4.4.2 Random walk Method

We use a graph using the SNOMED CT concepts relations as the edges for our graph and the concept's IDs as the vertices for our graph. Then we build random paths of length 10 from this graph with a predefined random probability of each edge. The generated random walks are used as input for the RNN model to generate word embeddings for the testing phrases from the benchmark datasets. According to the results in table 10 we found that this method outperforms both the SNOMED CT random factorization and the corpus-based embeddings.

	Random Walk Embeddings	
	Word2Vec model	RNN model
Pederson	0.471	0.61
Mayosrs	0.437	0.31
Umnsrs	0.53	0.41
Hliaoutakis	0.663	0.61
EHR-RelB	0.440	0.450

Table 12 Random Walk Embeddings. Pearson correlation coefficient between similarity scores from human judgments and those from word embeddings on four measurement datasets using random walk.

4.4.3 Utilize relation types and semantic types

In the results presented in the previous sections, only the graph structure was used to obtain embeddings. We used 5 semantic types and 5 relation types in order to read all the SNOMED

concepts and to test the effect of focusing on most common types and skipping the least used ones shown in Tables 12 and 13. This way the generated adjacency matrix dimension becomes lower and mode dense. We found that the correlation similarities are comparable with the corpus-based and random walk results. Then, we added another set of types, the results become better than before until we include all the semantic and relation types (62 sematic types and 27 relation types). We found that more medical semantic types add more information to the generated embeddings. Table 11 compares the correlation similarities for partial semantic and relation types and the complete selection for all semantic and relation types.

	Partial Semantic Types	All Semantic Types	Partial Relations Types	All Relation Types	SNOMED CT Random Walk
Pederson	0.4421	0.6033	0.5135	0.6221	0.471
Mayosrs	0.221	0.2926	0.4200	0.4369	0.437
Umnsrs	0.3101	0.3491	0.3081	0.3453	0.530
Hliaoutakis	0.4988	0.6290	0.3941	0.275	0.663

Table 13: Utilize relation types and semantic types. Pearson correlation coefficient between similarity scores from human judgments and those from word embeddings on four measurement datasets using random walk method. First column for most common semantic types. The second column uses all semantic types. Third column using selected concepts relations. Fourth column using all concepts relations.

'116680003'	isA
'363698007'	Finding site
'260686004'	Method
'116676008'	Associated Morphology

Table 14: Partial Relation Types.

‘288829000’	finding
‘64572001’	disorder
‘123037004’	body structure
‘71388002’	procedure

Table 15: Partial Semantic Types.

4.5 Obtaining Word and Clinical Term Embeddings from Clinical Concept Embeddings

4.5.1 SNOMED CT Term Embeddings Results

The methods described in the previous sections can obtain only clinical concept embeddings. However, for NLP applications one needs embeddings for words as well as clinical terms. This method has descent results for clinical terms from clinical corpus dataset. Results shown in Table 16 is competitive against the corpus based embeddings.

	SNOMED CT Embeddings using RNN	Corpus Based Embeddings			
		EHR	*MedLit*	*Glove*	*GoogleNews*
Pederson	0.81	0.632	0.569	0.403	0.357
Mayosrs	0.79	0.412	0.300	0.082	0.084
Umnsrs	0.67	0.440	0.404	0.177	0.154
Hliaoutakis	0.49	0.482	0.311	0.247	0.243
EHR-RelB	0.81	0.0	0.0	0.0	0.0

Table 16: Clinical Term embeddings using RNN model

4.5.2 Clinical Term Normalization Results

Table 15 shows the results for the clinical term normalization task on the MCN dataset [34] which was used in the n2c2 2019 shared task [53]. The first column shows results obtained by the full system. The second column shows the results when the patterns as described in the previous section are not used. The last column shows the results of only exact matching as a baseline for comparison. When the correct answer is not the top closest concept determined by the system, often it is one of the top closest concepts. Hence to gauge how far the correct answer is when the top answer is incorrect, the table also shows the results when the correct answer is within the top 2, 5 and 10 closest concepts.

	Clinical Term Normalization		
	Embeddings + Patterns + Exact	Embeddings + Exact	Exact only
Top 1	80.23	79.19	76.05
Top 2	82.37	82.23	78.05
Top 5	83.43	83.65	78.46
Top 10	83.78	83.97	78.46

Table 17: Results on the clinical term normalization task on the MCN benchmark dataset using the embeddings obtained from SNOMED CT using our method. The numbers are represent the accuracies (%) when the correct answer is in within the top 1, 2, 5, and 10 closest concepts according to the system.

Our system obtained 80.23% accuracy on this task. For comparison, the 33 teams that participated in the n2c2 2019 shared-task had obtained accuracies ranging from 51.85% to 85.26% with the top 10 teams obtaining accuracies ranging from 79.57% to 85.26% [34]. There was a large gap between the best (85.26%) and the second-best system (81.94%) system. These systems had used a variety of approaches and many of the top performing systems had specifically trained machine learning methods for the normalization task. In contrast, our system was not specifically trained for the normalization task, but it simply used embeddings learned from SNOMED CT to find the most similar concept. Yet our system performed competitively and would have secured 7th rank in this shared task based on accuracy. This shows that our method obtains embeddings for clinical terms which encode their concepts well enough that they can be used to normalize the clinical terms to their concepts.

From Table 11, one can observe that exact matching alone obtains 76.05% accuracy which is consistent with prior reporting [56]. With exact matching also sometimes, the correct answer is not the closest concept but the second-closest concept, this shows that a clinical term can be ambiguous and exactly match more than one concepts. The table also shows that the patterns helped in improving the accuracy from 79.19% to 80.23%. Given that the top-2 accuracy is

almost same with and without patterns, it shows that the patterns helped in determining the correct answer when it was within the top-2 closest concepts. When we used ClinicalBERT embeddings for normalization in the same way as we obtained results using our SNOMED CT embeddings, the accuracy was 78.3% without using patterns (worse than 79.19% accuracy obtained using SNOMED CT embeddings). With patterns, the accuracy improved to 80.16% (slightly worse than 80.23% obtained using SNOMED CT embeddings with patterns). This shows that the patterns are general and useful on this task when using corpus-based embeddings as well.

Besides quantitatively evaluating the embeddings obtained using SNOMED CT on two tasks, we qualitatively evaluated them and compared them with corpus-based embeddings. Table 18 shows five illustrative clinical terms, none of which is already present in UMLS, and the top 5 most similar clinical terms in UMLS found using the embeddings from SNOMED CT obtained by our method and found using embeddings from ClinicalBERT. The similarities between clinical terms were computed using cosine similarity between their embeddings. It can be observed that SNOMED CT embeddings found similar terms based on their clinical meanings, for example, for "broken thumb" it found "fracture of thumb" as most similar. In contrast, corpus-based embeddings found similar terms based on their linguistic usage, for example, for "broken thumb" it found "broken wrist" and "broken elbow" as most similar. Similar trend can be observed in the other examples too. This shows that embeddings obtained from SNOMED CT capture clinical semantics better than embeddings obtained from corpus-based methods.

Top 5 most similar terms using SNOMED CT embeddings	Top 5 most similar terms using ClinicalBERT embeddings
surgical removal of cancer	
excision of neoplasm; excision neoplasm malignant; excision of malignant neoplasm; excision tumor; excision tumors	surgical removal of prostate; surgical removal of gallbladder; surgical removal of impacted tooth; surgical removal of tooth; surgical removal of tonsil
pain in lower extremities	
pain in lower limb; pain in lower limb nos; pain in legs; pain in leg; limb pain leg; pain in unspecified lower leg	pain in upper extremities; pain in extremities; pain in bilateral lower legs; pain in bilateral upper arms; pain in upper arms
left toe injury	
injury of toe of right foot; injury of toe of left foot; open wound of right great toe; open wound of lesser toe of right foot; right toe contusion	left foot injury; left ankle injury; left thigh injury; left shoulder injury; right foot injury
pubic bone metastasis	
secondary malignant neoplasm of pubis; metastatic malignant neoplasm to pubis; metastatic malignant neoplasm to bone nos; bone neoplasm, malignant - pubis secondary; metastasis of malignant neoplasm to bone	dermal metastasis; adrenal gland metastasis; spleen metastasis; scrotal metastasis; axillary metastasis
broken thumb	
fracture of thumb; fracture thumb; fractured thumb; fracture of phalanges of thumb; fractures thumb	broken wrist; broken elbow; broken tooth; broken forearm; broken knee cap

Table 18 Qualitative comparison between the clinical term embeddings obtained from SNOMED CT using our method and clinical term embeddings obtained from ClinicalBERT. For each clinical term, the top 5 most similar terms in UMLS found using each type of embeddings are shown.

4.6 Combining Ontology-Based and Corpus-Based Embeddings

4.6.1 Concatenate Embeddings

By concatenating the Clinical BERT embeddings (dim = 768) with the SNOMED CT Random Walk embeddings (dim = 200). Table 16 shows a comparison between single embeddings and combined embeddings. We found that embedding combinations can improve the results. The first column is the correlation similarities for SNOMED CT Random Walk embeddings alone. The second column is the Clinical BERT Embeddings for the SNOMED CT concepts descriptions. The

third column shows the correlation coefficients for the concatenated embeddings. The results some improvement.

	SNOMED Random Walk	Clinical BERT Embeddings	SNOMED Random Walk + Clinical BERT embeddings
Pederson	0.61	0.58	0.62
Mayosrs	0.31	0.28	0.32
Umnsrs	0.41	0.28	0.41
Hliaoutakis	0.62	0.45	0.63
EHR-RelB	0.24	0.34	0.22

Table 19 Concatenate Embeddings. Correlation similarities for concatenated Clinical BERT embeddings with SNOMED CT Random Walk embeddings.

4.6.2 Merging Clinical corpus and SNOMED CT generated corpus

According to the following results in Table 17, the results of mering the trained clinical BERT embeddings with the original clinical BERT embeddings. We noticed an improvement in the results, that drives us to the result that multiple sources of embeddings increase the information.

	Pearson Correlation Similarity		Pearson Correlation Similarity using Trained Clinical BERT Embeddings
	SNOMED-CT generated corpus	Clinical corpus embeddings	SNOMED CT generated corpus + Clinical corpus
Pedersen's	0.18	0.08	0.48
Hliaoutakis's	0.20	0.01	0.17
MayoSRS	0.13	0.10	0.33
UMNSRS	0.10	0.23	0.22
EHR_RelB	0.156	0.23	0.05

Table 20 Merge clinical corpus and SNOMED generated corpus embeddings using RNN. Pearson correlation coefficient between similarity scores from human judgments and those from word embeddings on four measurement datasets.

In table 18, another way for correlation coefficient results by using the word representation model (word2vec)

Word representation Word2Vec Model			
	SNOMED Generated Corpus	Clinical Corpus	SNOMED Generated Corpus Combined with Clinical Corpus
Pederson	0.48	0.21	0.42
Mayosrs	0.17	0.086	0.13
Umnsrs	0.33	0.38	0.41
Hliaoutakis	0.22	0.2	0.23
EHR-RelB	0.05	0.05	0.01

Table 21 Merge Clinical and SNOMED corpus using Word2Vec. Pearson correlation coefficient between similarity scores from human judgments and those from word embeddings on four measurement datasets. A comparison table for word representation model (Word2Vec)

Table 19 shows an improvements in the correlation coefficient after training the clinical BERT with SNOMED CT generated corpus.

	Clinical BERT	**Clinical BERT trained on SNOMED generated corpus**
Medical Questions Corpus	0.13	0.41

Table 22 Pearson Correlation Similarities for Medical Questions Pairs.

4.6.3 Fine-Tuning BERT on SNOMED CT embeddings

It is commonly accepted that fine-tuning improves task performance. The fine-tuned models (along with the original models) are then used to generate contextualized representations. Our preliminary experiments showed that the commonly used 3-5 epochs of fine-tuning are insufficient for the smaller representations, such as BERT tiny, and they require more epochs. We fine-tuned all the representations for 10 epochs except BERT base, which we fine-tuned for the usual three epochs. Table 20 shows good results by fine-tuning SNOMED embeddings as target on a pretrained model on English language in a self-supervised fashion using a masked language modeling (MLM) objective.

Clinical BERT Fine Tuning		
	Fine-tuned Embeddings	Clinical Bert Embeddings
Pederson	0.588	0.08
Mayosrs	0.28	0.1
Umnsrs	0.28	0.23
Hliaoutakis	0.55	0
EHR-RelB	0.35	0.23

Table 23 Fine Tuning Clinical BERT. Pearson correlation coefficient between similarity scores from human judgments and those from word embeddings on four measurement datasets. Fine tuning a pretrained Clinical BERT using SNOMED CT embeddings description combination.

5 Future Work

We have many ideas that may improve the embeddings for the downstream NLP tasks. The suggested methods are considered as future work that can be continued work on leveraging the biomedical ontological knowledge to improve clinical term embeddings. The first one, is using more biomedical ontologies will increase the information given to embeddings which will positively affect the results. The other way that also could be another method which is add new semantic types and more relation types, since we notice that when we use all the semantic types and the relation types in SNOMED CT it increases the correlation similarities for medical term pairs. Using longer connections in the graph may help optimize the results which means considering multiple descendants for the medical concepts. It is like increasing the number of related concepts before and after the current concepts. The fourth idea is to use more grammatical words for the generated corpus and use more relations that will increase the size of the generated corpus. Lastly, Handling the linguistic morphology of the medical term and the edit pattern to solve the problem of singular and plural, polysemy and homonymy.

6 Conclusion

Traditionally, word embeddings are obtained from text corpora. In this research, we presented a novel method to obtain embeddings for clinical terms and words from the SNOMED CT ontology. The embeddings performed better than corpus-based embeddings on clinical term similarity task. They also performed competitively on clinical term normalization tasks. These results show that SNOMED CT is an alternate resource for obtaining clinical term embeddings and the presented method can successfully infuse ontological knowledge into embeddings. We also presented methods to combine the two sources of embeddings in order to incorporate linguistic as well as ontological knowledge into embeddings. The results on the evaluated tasks show that this helps improve the embeddings.

7 References

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

[2] SNOMED International, http://www.SNOMED.org/ (Accessed January 3, 2022).

[3] Zellig S Harris. Distributional structure. Word, 10(2-3):146–162, 1954.

[4] Charles E Osgood, George J Suci, and Percy H Tannenbaum. The measurement of meaning. Urbana: University of Illinois Press, 1957.

[5] L Wittgenstein. In gem anscombe. Philosophical investigations, 1953.

[6] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. Semantic similarity from natural language and ontology analysis. Synthesis Lectures on Human Language Technologies, 8(1):1–254, 2015.

[7] Noh, J., & Kavuluru, R. (2021). Improved biomedical word embeddings in the transformer era. Journal of Biomedical Informatics, 120, 103867.

[8] Agarwal, K., Eftimov, T., Addanki, R., Choudhury, S., Tamang, S., & Rallo, R. (2019). SNOMED2Vec: Random Walk and Poincare Embeddings of a Clinical Knowledge Base for Healthcare Analytics. arXiv preprint arXiv:1907.08650.

[9] Schulz, C., & Juric, D. (2020, April). Can Embeddings Adequately Represent Medical Terminology? New Large-Scale Medical Term Similarity Datasets Have the Answer! In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 87758782).

[10] Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., ... & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics, 87, 12-20.

[11] Saedi, C., Branco, A., Rodrigues, J., & Silva, J. (2018, July). Wordnet embeddings. In Proceedings of the third workshop on representation learning for NLP (pp. 122-131).

[12] Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. arXiv preprint arXiv:1411.4166.

[13] Yu, M., & Dredze, M. (2014, June). Improving lexical embeddings with semantic knowledge. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 545-550).

[14] Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2015). Nasari: a novel approach to a semantically-aware representation of items. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 567-577).

[15] Vendrov, I., Kiros, R., Fidler, S., & Urtasun, R. (2015). Order-embeddings of images and language. arXiv preprint arXiv:1511.06361.

[16] Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems, 30, 6338-6347.

[17] Rodrigues, J., Branco, R., Silva, J., Saedi, C., & Branco, A. (2018, July). Predicting brain activation with WordNet embeddings. In Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing (pp. 1-5).

[18] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.

[19] Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015, February). Learning entity and relation embeddings for knowledge graph completion. In Twenty-ninth AAAI conference on artificial intelligence.

[20] Nickel, M., Rosasco, L., & Poggio, T. (2016, March). Holographic embeddings of knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

[21] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137–1155, 2003.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[24] Hongyun Cai, Vincent W Zheng, and Kevin Chang. A comprehensive survey of graph embedding : problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 2018.

[25] Mark Newman. Networks: an introduction. Oxford university press, 2010.

[26] https://conservancy.umn.edu/handle/11299/196265 ((Accessed April 30, 2023))

[27] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pages 855–864. ACM, 2016.

[28] [28] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data. Machine Learning, 94(2):233–259, 2014.

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9 (Nov) : 2579–2605, 2008.

[30] Franz Baader, Diego Calvanese, Deborah McGuinness, Peter Patel-Schneider, Daniele Nardi (Eds.), The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, 2003.

[31] [31] Stephen T. Wu, Vinod C. Kaggal, Dmitriy Dligach, James J. Masanz, Pei Chen, Lee Becker, Wendy W. Chapman, Guergana K. Savova, Hongfang Liu, Christopher G. Chute (2013), A common type system for clinical natural language processing, J. Biomed. Semant. 4 (1), 1-12.

[32] Kate, R. J. (2020). Automatic full conversion of clinical terms into SNOMED CT concepts. Journal of Biomedical Informatics, 111, 103585.

[33] SNOMED CT Editorial Guide, https://confluence.ihtsdotools.org/display/DOCEG (Accessed April 30, 2023).

[34] Yen-Fu Luo, Weiyi Sun, Anna Rumshisky, MCN: A comprehensive corpus for medical concept normalization, J. Biomed. Inform. (2019) 103132.

[35] SNOMED CT Editorial Guide, https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=26837115 (Accessed April 30, 2023)

[36] Salawa, M., Branco, A., Branco, R., Rodrigues, J., & Saedi, C. (2019, September). Whom to learn from? graph-vs. text-based word embeddings. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 10411051).

[37] Hliaoutakis, A. (2005). Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline. Master's thesis.

[38] Pakhomov, S. V., Pedersen, T., McInnes, B., Melton, G. B., Ruggieri, A., & Chute, C. G. (2011). Towards a framework for developing semantic relatedness reference standards. Journal of biomedical informatics, 44(2), 251-265.

[39] Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics, 40(3), 288-299.

[40] Kate, R. J. (2021). Clinical term normalization using learned edit patterns and sub-concept matching: system development and evaluation. JMIR Medical Informatics, 9(1), e23104.

[41] Schulz, C., Levy-Kramer, J., Van Assel, C., Kepes, M., & Hammerla, N. (2020). Biomedical Concept Relatedness--A large EHR-based benchmark. arXiv preprint arXiv:2010.16218.

[42] https://confluence.ihtsdotools.org/display/DOCSTART/4.+SNOMED+CT+Basics (Accessed April 30, 2023).

[43] He, Y., Chen, J., Antonyrajah, D., & Horrocks, I. (2021). BERTMap: A BERT-based Ontology Alignment System. arXiv preprint arXiv:2112.02682.

[44] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

[45] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

[46] Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural language processing (NLP) clinical challenges (n2c2)/open health NLP (OHNLP) shared task on clinical

concept normalization for clinical records. J Am Med Inform Assoc 2020 Oct 1;27(10):1529-1537.

[47] Noh J, Kavuluru R. Improved biomedical word embeddings in the transformer era. Journal of Biomedical Informatics. 2021 Aug 1;120:103867.

[48] Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY. A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems. 2020 Mar 24;32(1):4-24.

[49] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data. 2019 May 10;6(1):52.

[50] Xu C, Bai Y, Bian J, Gao B, Wang G, Liu X, Liu TY. RC-NET: A general framework for incorporating knowledge into word representations. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management 2014 Nov 3 (pp. 1219-1228).

[51] Alawad M, Hasan SS, Christian JB, Tourassi G. Retrofitting word embeddings with the UMLS metathesaurus for clinical information extraction. In2018 IEEE International Conference on Big Data (Big Data) 2018 Dec 10 (pp. 2838-2846). IEEE.

[52] Pattisapu N, Patil S, Palshikar G, Varma V. Medical concept normalization by encoding target knowledge. In Machine Learning for Health Workshop 2020 Apr 30 (pp. 246-259). PMLR.

[53] Luo YF, Henry S, Wang Y, Shen F, Uzuner O, Rumshisky A. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. Journal of the American Medical Informatics Association. 2020 Oct;27(10):1529-e1.

[54] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004 Jan 1;32(suppl_1):D267-70.

[55] Kate RJ. Normalizing clinical terms using learned edit distance patterns. Journal of the American Medical Informatics Association. 2016 Mar 1;23(2):380-6.

[56] Kate RJ. Clinical term normalization using learned edit patterns and sub concept matching: system development and evaluation. JMIR Medical Informatics. 2021 Jan 14;9(1):e23104.

[57] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv:1802.05365 [cs]. ArXiv: 1802.05365.

[58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]. ArXiv:1810.04805.

[59] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv:1901.08746 [cs]. ArXiv: 1901.08746.

[60] Kaung Khin, Philipp Burckhardt, and Rema Padman. 2018. A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation. arXiv:1810.01570 [cs]. ArXiv: 1810.01570.

[61] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing Clinical Concept Extraction with Contextual Embedding. arXiv:1902.08691 [cs]. ArXiv: 1902.08691.

[62] Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. Proceedings of LBM. 2013 Dec 12:39-44.

[63] Kosmopoulos A, Androutsopoulos I, Paliouras G. Biomedical semantic indexing using dense word vectors in bioasq. J BioMed Semant Suppl BioMedl Inf Retr. 2015;3410:959136040-1510456246.

[64] Chiu B, Crichton G, Korhonen A, Pyysalo S. How to train good word embeddings for biomedical NLP. InProceedings of the 15th workshop on biomedical natural language processing 2016 Aug (pp. 166-174).

[65] McDonald R, Brokos GI, Androutsopoulos I. Deep relevance ranking using enhanced document-query interactions. arXiv preprint arXiv:1809.01682. 2018 Sep 5.

[66] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data. 2019 May 10;6(1):52.

[67] Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. In2019 IEEE International Conference on Healthcare Informatics (ICHI) 2019 Jun 10 (pp. 1-5). IEEE.

[68] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019 Apr 6.

[69] McCreery, Clara H., Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. "Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs." In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3458-3465. 2020.