Journal of Biomedical Informatics, Volume 149, January 2024,104560.

Obtaining Clinical Term Embeddings from SNOMED CT Ontology

Fuad Abu Zahra and Rohit J. Kate*

Department of Computer Science

University of Wisconsin-Milwaukee

Milwaukee, WI, USA

ABSTRACT

Clinical term embeddings are traditionally obtained using corpus-based methods, however, these methods cannot incorporate knowledge about clinical terms which is already present in medical ontologies. On the other hand, graph-based methods can obtain embeddings of clinical concepts from ontologies, but they cannot obtain embeddings for clinical terms and words. In this paper, a novel method is presented to obtain embeddings for clinical terms and words from the SNOMED CT ontology. The method first obtains embeddings of clinical concepts from SNOMED CT using a graph-based method. Next, these concept embeddings are used as targets to train a deep learning model to map clinical terms to concepts embeddings. The learned model then provides embeddings for clinical terms and words as well as maps novel clinical terms to their embeddings. The embeddings obtained using the method out-performed corpus-based embeddings on the task of predicting clinical term similarity on five benchmark datasets. On the clinical term normalization task, using these embeddings simply as a means of computing similarity between clinical terms obtained accuracy which was competitive to methods trained specifically for this task. Both corpus-based and ontology-based embeddings have a limitation that they tend to learn similar embeddings for opposite or analogous terms. To counter this, we also introduce a method to automatically learn patterns that indicate when two clinical terms represent the same concept and when they represent different concepts. Supplementing the normalization process with these patterns showed improvement. Although clinical term embeddings obtained from SNOMED CT incorporate ontological knowledge which is missed by corpus-based embeddings, they do not incorporate linguistic knowledge which is needed for sentence-based tasks. Hence combining ontology-based embeddings with corpus-based embeddings is an avenue for future work.

Keywords: clinical terms; SNOMED CT; embeddings; ontology

* Corresponding author: katerj@uwm.edu

1. Introduction

Deep learning based methods have shown success on several natural language processing (NLP) tasks, including in the clinical domain [1]. A critical component of all deep learning based NLP methods is the representation of words in numerical vector forms, also known as word embeddings. Given that neural networks can only take input in numerical form, word embeddings provide a suitable mechanism to give words, which are otherwise symbolic, as input to neural networks. Additionally, from machine learning perspective, they also provide a way to generalize from words seen during training to those not seen during training by leveraging the fact that words with similar meanings have similar word embeddings. Word embeddings are commonly obtained using corpus-based methods [2] which work on the basic premise that words found in similar contexts would have similar meanings and hence should have similar word embeddings. Although this is a reasonable premise, corpus-based methods expect all words to occur frequently enough in the corpus so that their embeddings can be suitably learned. However, this is not always true, especially in the clinical domain where the names of many diseases or medications may not occur frequently in a corpus. For example, consider disease names “pneumonia” and “pneumoconiosis” which are both inflammatory disorders of lungs and hence have similar meanings. However, for a corpus-based method to learn similar word embeddings for them, these words will need to be present in the corpus in similar contexts multiple times, which may not happen in a clinical corpus. In addition, embedding for each synonym of the disease will have to be learned independently. Given that there are more than a million clinical terms, it is not surprising that corpus-based embeddings were not found to do well on clinical term similarity prediction task [3].

Ontologies encode knowledge of a domain in the form of a graph, with concepts as nodes and relations between them as edges [4]. Medical ontologies, such as SNOMED CT [5], directly encode semantic properties of medical concepts in the graph. For example, the concepts of “pneumonia” and “pneumoconiosis” are both linked by “is-a” relation to the concept of “disorder of respiratory structure”, are linked by “finding site” relation to the concept “lung structure”, and are linked by “associated morphology” relation to the concept of “inflammation”. Given that they share multiple relations with other concepts, it can be explicitly and directly inferred from SNOMED CT that the concepts of “pneumonia” and “pneumoconiosis” are similar. In contrast, this can only be learned implicitly and indirectly by corpus-based methods from their contexts that too if they occur frequently enough, as pointed out earlier. Hence knowledge from ontologies could be used as an alternate resource for learning word embeddings.

In the general domain, WordNet ontology has been used to learn word embeddings using graph-based methods [6]. However, unlike WordNet in which words themselves are the nodes of the graph, in medical ontologies medical concepts are the nodes of the graph. A medical concept (typically denoted by an identifier in an ontology) may be associated with multiple terms, each with multiple words. For example, there is a concept of viral meningitis (id=58170007) in SNOMED CT, with associated clinical terms (also known as *descriptions* in SNOMED CT) "viral meningitis", "abacterial meningitis" and "aseptic meningitis, viral". While a graph-based method will obtain embedding for the concept 58170007, it will not obtain embeddings for words such as "meningitis", "viral", etc. It will also not give embeddings for previously unseen clinical terms even though they may be composed of previously seen words, such as "bacterial meningitis".

In this paper, we present a novel method to obtain clinical term and word embeddings from SNOMED CT ontology. After obtaining embeddings for concepts using a graph-based method, a deep learning network is trained to map clinical terms to these embeddings. In this process, the network learns the embeddings of clinical terms and words, including their synonyms, as well as learns to obtain embeddings for previously unseen clinical terms. To the best of our knowledge, this is the first method that obtains clinical term embeddings from clinical concept embeddings. Using standard benchmark datasets, the method was evaluated on clinical term similarity prediction task and on clinical term normalization task. Both corpus-based and ontology-based embeddings suffer from the limitation that they tend to obtain similar embeddings for terms with opposite or analogous meaning, for example, "left kidney" and "right kidney". Although, they seem similar, "left kidney" and "right kidney" clinically mean very different things and hence should not be treated as similar. To counter this limitation of embeddings, we also introduce a method to automatically learn patterns from UMLS [7] which indicate whether two clinical terms would have same meaning or not. Not only these patterns showed improvement in normalization performance for both corpus-based and ontology-based embeddings, but they could also be used as a resource to further improve clinical term embeddings in future.

2. Related Work

Although there has been a lot of work in obtaining embeddings using corpus-based methods in the clinical domain [2], there has been relatively less work in obtaining embeddings from biomedical ontologies. Some researchers obtained embeddings of concepts in UMLS [8,9], but we note that although UMLS Metathesaurus combines clinical concepts from multiple sources, it does not encode meanings of concepts in terms of their relations with other concepts as is done in SNOMED CT which is an ontology based on the description logic framework [10]. Hence SNOMED CT is a better candidate for learning

meaning-based embeddings. Agrawal et al. [11] obtained embeddings from SNOMED CT using graph-based methods, however, they obtained embeddings of clinical concepts, not terms or words. Consequently, they could evaluate those embeddings only on concept related tasks, such as, classifying the relation between two concepts, and not on tasks related to clinical terms. Using a method similar to OWL2Vec* [12] from the general domain, Castell-Díaz et al. [13] recently obtained knowledge graph embeddings for clinical concepts and clinical terms together from SNOMED CT. Our method is very different from theirs because we first obtain clinical concept embeddings using a graph-based method and then train a deep learning model to map clinical terms to these embeddings. An advantage of our method is that the trained model can directly give embedding of a multi-token clinical term, in contrast, their method requires adding and averaging the embeddings of the individual tokens. Furthermore, their embeddings were aimed specifically for the task of creating new SNOMED CT post-coordinated concepts from clinical terms [14] and were evaluated only for that task.

In past, corpus-based embeddings have been retrofitted to ontological relations in the general domain [15,16] as well as in the clinical domain [17,18]. An approach to transform corpus-based embeddings to minimize their cosine similarity with graph-based embeddings was presented in [19]. However, these approaches only indirectly use ontological knowledge to influence corpus-based embeddings, they do not directly learn embeddings from ontology. Noh and Kavuluru [20] presented a method where MeSH [21] concept codes are directly inserted into a corpus in place of the clinical terms and then a corpus-based method is used to jointly learn embeddings for both the words and the concepts. Although this approach is a good way to learn embeddings for concepts using a corpus, it does not leverage ontological relations.

3. Materials and Methods

3.1. Obtaining Clinical Term Embeddings

We chose to use SNOMED CT as the ontology for obtaining embeddings because it is the most comprehensive medical ontology, and unlike other resources such as UMLS [7] and MeSH [21], SNOMED CT defines concepts in terms of their relations with other concepts thus making it more conducive for learning meaning-based embeddings. Given the SNOMED CT ontological graph with the concepts as nodes and the relations between them as the edges, our method first uses the random walk method to obtain embeddings of the concepts [11]. In this method, the graph is randomly traversed from one node to another and these traversed paths are collected as if they were “sentences” with the nodes being the “words”, thus forming a “corpus”. Next, a corpus-based method, skip-gram [22], is applied on this synthetic “corpus” to obtain embeddings for the nodes. The dimension of embeddings was 200 which was found to work well in the previous work [11]. The premise behind this method is that similar

concepts will have similar relations to other concepts which will act as similar "contexts" and hence will lead to similar embeddings.

As mentioned in the Introduction, this method can only obtain embeddings of the clinical concepts which form the nodes of the graph, but we want embeddings of clinical terms and words, as well as a mechanism to obtain embeddings of previously unseen clinical terms. To accomplish this, we developed a novel method. In this method, a deep learning model is trained which takes clinical terms as input and the corresponding concept embeddings as targets. Figure 1 shows the network along with an illustrative example. Given that clinical terms could be of multiple words and vary in length, we used a recurrent neural network model [23]. We did not find it necessary to use attention-based models [24] because clinical terms are not too long to need long-distance attentions. Figure 1 shows the recurrent network unrolled over time for the clinical term of length three. The first layer is the embedding layer which is where the embeddings for words get learned to suit the task. This is followed by two GRU layers [25] and a fully connected layer. The target is the concept embedding corresponding to the input term as obtained from the graph-based method. Thus, the network learns to map clinical terms to their concept embeddings. The network was implemented using the Keras deep learning package [26].

To create training examples for this model, every concept in SNOMED CT is used as target and is paired with its every description (i.e., its fully specified name and every synonym) in SNOMED CT as well as its every synonym from UMLS as input. Figure 1 shows the embedding of the clinical concept upper abdominal pain (SNOMED CT id=83132003) which was obtained earlier using the graph-based method being used as the target with its description "upper abdominal pain" as input. As part of the training process of mapping the clinical term to the target concept embedding, the network will learn the embeddings for the words "upper", "abdominal" and "pain" in the embedding layer. For training, 10 epochs were found sufficient for convergence.

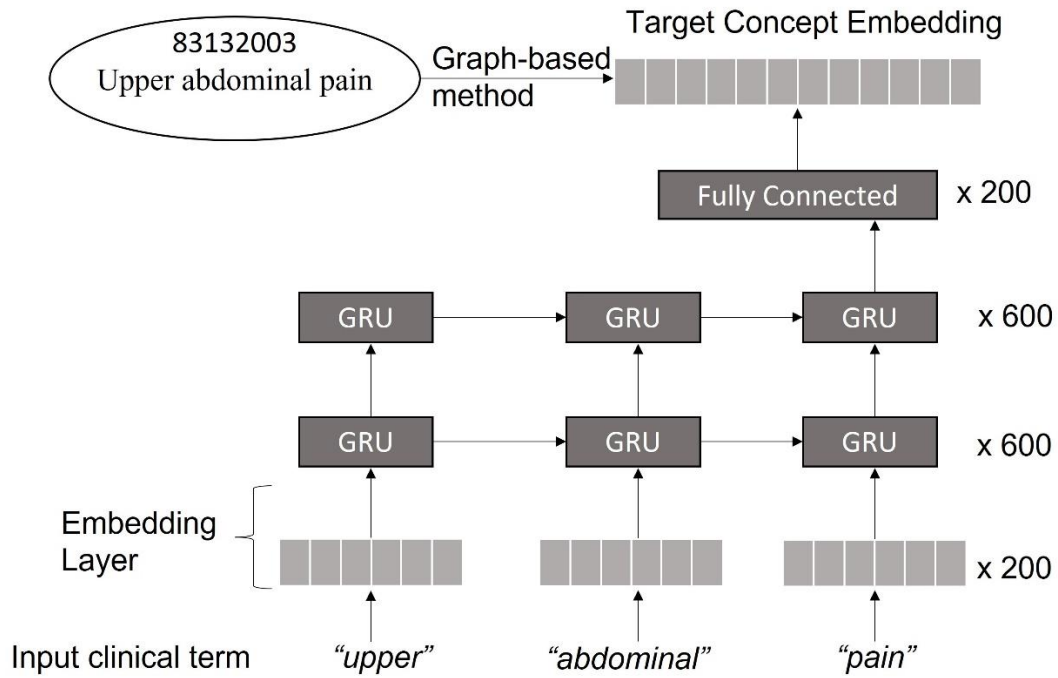

Figure 1. Deep learning network that maps clinical terms to their concept embeddings obtained using a graph-based method and in this process learns the embeddings for clinical terms and words through the embedding layer. The figure shows the recurrent neural network unrolled over time for a three token input.

Not only this network can obtain embeddings for words, but it can also give embeddings for clinical terms previously unseen by the network. For example, after it has been trained, if the network shown in Figure 1 is given input "acute upper abdominal pain", it will give its concept embedding even though the term or the concept is not present in SNOMED CT or UMLS. We note that it is very common to encounter clinical terms not already present in terminologies, for example, one study estimated that 19.75% of clinical terms mentioned in text did not have their concepts present in SNOMED CT [27]. There are two main reasons for this. First, concepts can be compositionally created in medicine from other concepts and no terminology can exhaustively list all possible concepts and their corresponding terms. Second, variability in natural languages allows one to express a clinical concept in multiple ways. We also found in our experiments with clinical term similarity and clinical term normalization tasks that clinical terms often do not match exactly in clinical terminologies, and in these situations the ability of our method to provide embeddings for previously unseen clinical terms is crucial.

Our method ends up learning embeddings for all the words used in the clinical terms in UMLS which offers a wide coverage of clinical words. But in case it encounters an unknown word in a clinical term, it uses a random embedding for the word to compute embedding for the clinical term. We found that random embeddings worked better than using a designated unknown word or using all zeros as an

embedding because otherwise the method would incorrectly regard two unknown words to be same during similarity matching. In future, our method could be used to learn character-based embeddings, or it could employ subword tokenization as used by the BERT system in order to better handle unknown words.

3.2. Clinical Term Similarity

The first task on which we evaluated clinical term embeddings obtained from SNOMED CT is the clinical term similarity task. In this task, given two clinical terms, a method has to predict similarity between them which is then compared against the expert-judged similarity score. We used five benchmark datasets for this task whose details are depicted in Table 1. Each dataset consists of a list of clinical term pairs along with their expert-judged similarity scores. The first four datasets have been widely used in the past [28], more recently, a fifth dataset "EHR-RelB" was introduced which is much larger and consists of longer clinical terms [3].

Dataset	Number of clinical term pairs	Average clinical term length
Pedersen's [29]	29	1.57
Hliaoutakis's [30]	35	1.69
MayoSRS [31]	101	1.55
UMNSRS [32]	566	1.02
EHR-RelB [3]	3630	3.04

Table 1. Benchmark datasets used for evaluation for the clinical term similarity task.

To obtain embedding of a clinical term, it is given as input to our trained model described in the previous section, which then outputs its embedding. Given a pair of clinical terms, similarity between them is computed as the cosine similarity between their embeddings. 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. We compare the performance of SNOMED CT embeddings with corpus-based embeddings on this task.

3.3. Clinical Term Normalization

The second task on which we evaluated clinical term embeddings obtained from SNOMED CT is the clinical term normalization task [33]. 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 [7]. For example, given a clinical term "pain in abdomen", it should be normalized to the concept which has the UMLS concept unique identifier (CUI) of C0000737. This task is challenging because a clinical term does not always exactly match the clinical terms listed in a terminology due to the variability allowed in natural languages such as English. For example, the clinical term "pain in abdomen" does not exactly match any clinical term in UMLS even though UMLS has several clinical terms listed for that concept, such as

"abdominal pain", "pain in stomach", "gut pain" and "bellyache". For our experiments, we used the benchmark MCN dataset [27] which has been used extensively for evaluating normalization methods [33]. 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 CUIs. We also evaluated on the ShARe/CLEF eHealth 2013 dataset [34] for its normalization subtask. This dataset contains 5,816 clinical terms for training and 5,351 clinical terms for testing. Unlike MCN dataset, these clinical terms are restricted to only the disease/disorder semantic type. In this entire dataset, 30.3% of clinical terms are "CUI-less".

3.3.1. Normalization Using Embeddings

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 the average similarity between the given clinical term and all the clinical terms in UMLS corresponding to each of those concepts is computed and the concept with the highest similarity is given as the output.

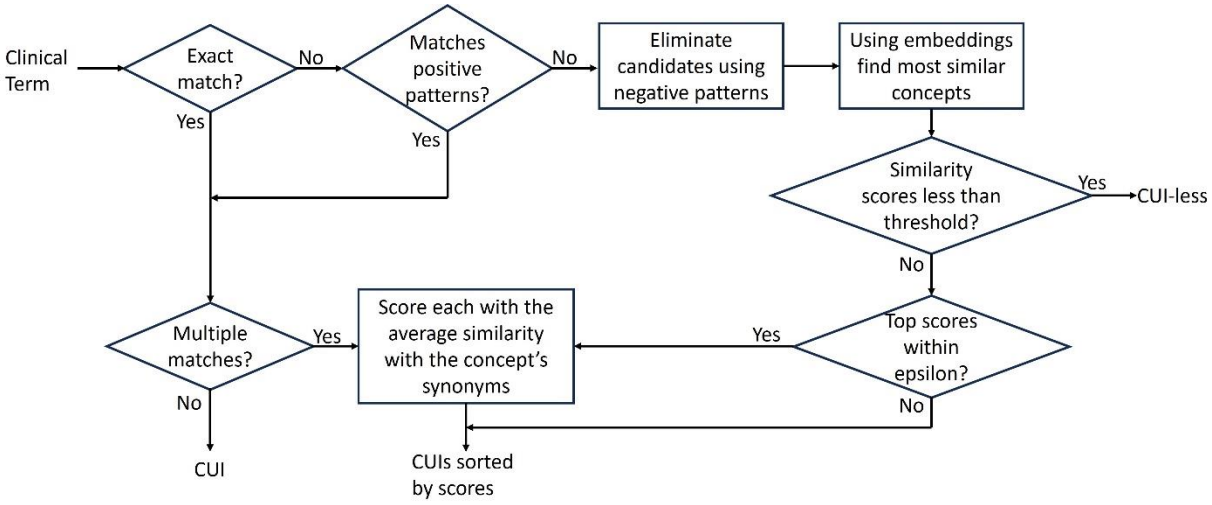

Figure 2. Flowchart depicting the clinical term normalization process.

If the clinical term does not match exactly either in UMLS or in the training examples, 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. However, if the difference between the similarity of the top concepts is too close (less than epsilon=0.001) then the average similarity with all the clinical terms in UMLS corresponding to those top concepts are computed and the concept with the highest similarity is given as the output (analogous to

how it is done when there are multiple exact matches). If no clinical term is found in UMLS with similarity more than threshold=0.9 then "CUI-less" is given as the output. The entire normalization process is shown as a flowchart in Figure 2 (the use of patterns is explained in the next subsection).

For efficiency, the embeddings of all the clinical terms in UMLS are pre-computed using our 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 (giving an accuracy gain of around 1% absolute on the normalization task), 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. This is the similarity measure which is used in the method described above.

We observed a limitation of the embeddings obtained from SNOMED CT which is also a limitation of the embeddings obtained using corpus-based methods. The model learns very similar embeddings for terms 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". For example, "primary tumor" should not be normalized to "secondary tumor". However, the learned embeddings will tend to do so because they would learn similar embeddings for "primary" and "secondary". This happens because the clinical terms with opposite or analogous meanings will have their concepts in very similar positions in the ontological graph. For example, the concepts of "left kidney" and "right kidney" will be related to the same other concepts with the same relations, except for "laterality". As a result, our model tends to learn very similar embeddings for such terms even though they have clinically very different meanings. Corpus-based embeddings also suffer from this limitation because words with opposite or analogous meanings are often found in similar contexts in text and hence corpus-based methods also learn similar embeddings for them.

Another limitation we observed was that the model sometimes would learn different embeddings for terms 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". 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 in the training corpus.

To counter the above limitations, we supplemented our normalization method with some patterns which were automatically learned from UMLS as described in the next subsection. We later include results of an ablation study that shows how much they contributed to the normalization task.

3.3.2. Supplementing Normalization Method with Patterns

Each of our patterns 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 two clinical terms "primary neoplasm" and "secondary neoplasm", the pattern derived from them will be "primary | secondary", where the two parts of the pattern are shown separated by "|". The pseudocode to derive a pattern from two terms is shown as the first function in Figure 3. Words(T) represents the set of words in term T on which set operations are then applied.

```
derivePattern(Term1, Term2) :
    # Takes two terms and returns their derived pattern.
    firstPart = Words(Term1) – Words(Term2)
    secondPart = Words(Term2) – Words(Term1)
    return (firstPart | secondPart)

patternMatches((P1|P2), Term1, Term2) :
    # Returns if pattern (P1|P2) matches pair of terms Term1 and Term2
    (R1|R2) = derivePattern(Term1, Term2)
    return (R1==P1 and R2==P2) or (R2==P1 and R1==P2)

learnPatterns(UMLS) :
    # Learns positive and negative patterns from UMLS.
     patterns = Empty
     for every T1 and T2 in UMLS :
        # for every two terms in UMLS
        if |Words(T1) ∩ Words(T2)| >
                |Words(T1) U Words(T2)|/2 :
            # at least half words are common
            P = derivePattern(T1, T2)
            if length(P) < 5 :
                # combined length should be less than 5
                if P is not in patterns :
                    P.pos = 0   P.neg = 0
                    Include P in patterns
                if Concept(T1) == Concept(T2) :
                    # Terms represent the same concept
                    patterns[P].pos += 1 # increment positive count
                else :
                    patterns[P].neg += 1  # increment negative count
     posPatterns = All P in patterns such that P.pos > 5 and P.pos > 10*P.neg
     negPatterns = All P in patterns such that P.neg > 5 and P.neg > 10*P.pos
     return posPatterns, negPatterns
```

Figure 3. Pseudo-code for the methods to derive, match, and learn patterns from UMLS. Words(T) represents the set of words in term T. The pseudo-code uses set operations on the set of words. Concept(T) represents the concept in UMLS corresponding to the term T.

There are two types of patterns – *positive patterns* and *negative patterns* (we describe it later how we obtained them). If a pattern is positive, then it means that replacing words from its one part in a clinical term with the words from its second part does not change the meaning of the clinical term. For example, the pattern "bilateral | both" is positive because replacing the words "bilateral" with "both" does not change the meaning of a clinical term (for example, "edema of bilateral lower extremities" and "edema of both lower extremities"). In contrast, if a pattern is negative then it means that replacing words from its one part in a clinical term with the words from its second part changes the meaning of the clinical term. The pattern "primary | secondary" is a negative pattern because replacing "primary" with "secondary" in a clinical term changes its meaning (for example, "primary tumor" and "secondary tumor"). The two parts of the pattern are considered inter-changeable when being applied (in other words the pattern "secondary | primary" is equivalent to the pattern "primary | secondary").

One of the parts of a pattern could also be empty which would capture whether presence of extra words changes the meaning of a clinical term or not. For example, the positive pattern "nos | " indicates that presence of "nos" (meaning "not specified") does not change the meaning when added to a clinical term (or removed from it), whereas the negative pattern "infected | " indicates that presence of "infected" changes the meaning when added to a clinical term (or removed from it). A part of a pattern can also have multiple words, for example, "bilateral | left and right" which is a positive pattern indicating that "bilateral" can be replaced by "left and right" in a clinical term without changing its meaning. The words within a part are treated like bag of words (that is, their order is ignored). We found that this made the patterns more general, thus improving the performance, while losing the word order rarely created a problem. For example, the same pattern will also match "right and left" in a clinical term, while an ungrammatical word order, such as "left right and", is unlikely to be present in a clinical term. We note that these patterns are different from patterns from our past work [35,36], 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.

A pattern can be matched against a pair of clinical terms to determine whether the two clinical terms mean the same thing (in case a positive pattern matches) or cannot mean the thing (in case a negative pattern matches). The pseudocode to match a pattern against a pair of clinical terms is shown as the second function in Figure 3. If the pattern derived from the pair of clinical terms is same as the given pattern, then we say that the pattern matched the pair of the clinical terms. For example, the pattern "bilateral | left and right" will match the pair of clinical terms "ultrasonography of bilateral kidneys" and

"ultrasonography of left and right kidneys". This is because in the pattern derived from this pair of clinical terms, the first part will be "bilateral" which is the only word in the first term which is not in the second term, while the second part will be "left and right" which are the only words in the second term which are not in the first term. Thus, the pattern "bilateral | left and right" matches these two clinical terms. Given that it is a positive pattern, it can be concluded that the two terms mean the same thing or map to the same concept.

We used the following simple method to automatically learn positive and negative patterns from UMLS. Its pseudocode is shown as the third function in the Figure 3. The method first derives patterns and then determines whether they are positive patterns or negative patterns or neither. The patterns are first derived by considering every two clinical terms in UMLS. As mentioned earlier, the words which are present in the first clinical term but not in the second clinical term become the first part of the pattern, and the vice-versa become the second part of the pattern. To make this process efficient and to also find more useful patterns, only those pairs 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 are restricted to have the combined length of the two parts to be less than five.

Once a pattern has been derived, its number of *positive matches* and *negative matches* are counted. If a pattern matches two clinical terms in UMLS which share the same concept then it is considered a positive match, but if they do not share the same concept then it is considered a negative match implying that the two clinical terms mean different things. In the shown pseudocode, the counting is done along with deriving the patterns in the same loop that iterates over every two clinical terms in UMLS. A newly derived pattern is included in the set of patterns if it is not already present. Whether the pattern is already present or newly included, its count for either positive matches or negative matches is incremented by one based on whether the two clinical terms (it was just derived from) share the same concept or not. Ambiguous clinical terms (that are associated with more than one concept) are not included in this learning process.

After the process of deriving all the patterns and counting their number of positive and negative matches, the patterns are identified as either positive patterns or negative patterns or neither of them. We call the patterns as *positive patterns* if they have more than 5 positive matches and 10 times more positive matches than negative matches. Similarly, we call the patterns *negative patterns* if they have more than 5 negative matches and 10 times more negative matches than positive matches. The patterns which are neither positive patterns nor negative patterns are simply dropped. A minimum of 5 matches are required

so that the patterns are not too rare, and 10 times matches are required so that the patterns are at least 90% accurate (a goal for the normalization task). The method learned total 11,236 positive patterns and 6,078,532 negative patterns from UMLS. Table 2 shows a few examples of positive patterns and negative patterns learned from UMLS along with their number of positive and negative matches.

These learned patterns are supplemented in the normalization method in a simple way as follows which is also shown in the flowchart of Figure 2. The negative patterns are used to eliminate candidate concepts. If a negative pattern matches the clinical term to be normalized paired with a candidate clinical term in UMLS, then the concept corresponding to that clinical term can never be the output. This is because a match with a negative pattern indicates that the two clinical terms have different meanings. The negative patterns thus mitigate the limitation that similar embeddings may be learned for clinical terms with opposite or analogous meanings. For example, the negative pattern "primary | secondary" will not allow "primary tumor" to be normalized to "secondary tumor" even though the two terms may have very similar embeddings. The positive patterns are used to find clinical terms which are treated like exact matches. If a positive pattern matches the clinical term to be normalized paired with a clinical term in UMLS, then it is treated as if the two terms matched exactly. This is because a match with a positive pattern indicates that the two clinical terms have the same meaning. The positive patterns thus mitigate the limitation that different embeddings may be learned for terms with similar meanings. For example, the positive pattern "bilateral | left and right" will normalize "left and right kidneys" to "bilateral kidneys" even though the two terms may have very different embeddings. The rest of the normalization process proceeds as described in the previous subsection.

Our trained model and the learned patterns are available through the website: https://sites.uwm.edu/katerj/JBI2023.

Example Patterns	Positive Matches	Negative Matches
Positive Patterns		
metastatic to | secondary of	480	0
assay | measurement of	231	0
hepatic | liver	206	0
subcutaneous injection | percutaneous	196	0
ultrasound scan | ultrasonography	193	0
k+ | pot	185	0
bilateral extremities | both limbs	31	0
Negative Patterns		
benign | malignant	0	2770
anterior | posterior	0	2670
lumbar | thoracic	0	1622
artery | vein	0	1245
bilateral eyes | left eye	0	540
moderate |	0	342
right structure of |	0	243

Table 2. A few examples of positive and negative patterns automatically learned from UMLS along with their positive and negative matches in UMLS. The two parts of a pattern are shown separated by "|", a part could be also empty. Replacing words from one part of a pattern with words from the other part does not change meaning of a clinical term for a positive pattern but changes meaning for a negative pattern.

4. Results and Discussion

4.1. Clinical Term Similarity

Table 3 shows the results of the clinical term similarity task on the four benchmark datasets comparing the embeddings obtained from SNOMED CT using our method with the embeddings obtained using a few corpus-based methods as reported in Wang et al. 2018 [28]. We added the results we obtained using embeddings from ClinicalBERT [37]. The numbers in the table are Pearson correlation coefficient between the similarity scores obtained using the embeddings and the expert-judged similarity scores. Among the corpus-based embeddings, "EHR" embeddings were obtained using a clinical corpus from electronic health records, "MedLit" embeddings were obtained using a corpus of medical literature, and the embeddings "Glove" and "Google News" were obtained using general domain corpora [28].

SNOMED CT based embeddings performed consistently better than the corpus-based embeddings on each dataset. This shows that the knowledge about meanings of clinical terms that is indicative of similarity between them can be better gleaned from SNOMED CT than from text corpora.

Dataset	SNOMED CT Embeddings	Corpus-based Embeddings				
		EHR	MedLit	Glove	Google News	ClinicalBERT
Pedersen's	**0.81**	0.63	0.57	0.40	0.36	0.08
Hliaoutakis's	**0.79**	0.48	0.31	0.25	0.24	0.00
MayoSRS	**0.67**	0.41	0.30	0.08	0.08	0.10
UMNSRS	**0.49**	0.44	0.40	0.18	0.15	0.23

Table 3. Results on clinical term similarity benchmark datasets comparing embeddings obtained from SNOMED CT using our method with a few corpus-based embeddings – four of them as reported in Wang et al. 2018 [28] and ClinicalBERT. The numbers are Pearson's correlation coefficient between the similarity scores obtained using the embeddings and the expert-judged similarity scores.

Table 4 shows the results on the larger and more complex clinical term similarity dataset introduced by Schultz et al. 2020 [3]. For comparison, results from several other open-source embeddings are also shown as reported in Schultz et. al. 2020 [38] in terms of Spearman's ranked correlation coefficient. The sources of these embeddings were as follows: PMC (PubMed Central), PM (PubMed), PP (both) and PPW (both plus Wikipedia) – [39], ASQ (BioASQ challenge dataset) – [40], LTL2 (Language Technology Lab, window size 2) and LTL30 (window size 30) – [41], AUEB2 (Athens University of Economics and Business, dimensionality 200) and AUEB4 (dimensionality 400)– [42], extr (extrinsic tasks) and intr (intrinsic tasks) – [17], and MIM (MIMIC) and MIM M (MIMIC and its model) – [43]. We added the result we obtained using embeddings from ClinicalBERT [37]. It can be observed that SNOMED CT based embeddings did better than other embeddings except one (ASQ) which was slightly better than it. Thus the results on this larger dataset further confirms that embeddings that encode knowledge about clinical term meanings can be obtained from SNOMED CT. Embeddings from ClinicalBERT embeddings did not do well on this dataset either. Their performance might have been limited because they are contextualized embeddings while this task does not provide contexts for the clinical terms. Our findings are consistent with previous findings that corpus-based embeddings, including contextual embeddings, do not capture semantics of clinical terms well [3].

SNOMED CT	PMC	PM	PP	PPW	ASQ	LTL2	LTL30	AUEB2	AUEB4	extr	intr	MIM	MIM M	Clinical BERT
0.45	0.40	0.44	0.42	0.41	**0.47**	0.36	0.41	0.40	0.40	0.35	0.37	0.33	0.33	0.23

Table 4. Results comparing embeddings obtained from SNOMED CT using our method with a few corpus-based embeddings (as reported in Schultz et al. 2020 [38]) and ClinicalBERT on the EHR-RelB benchmark dataset. The numbers are Spearman's ranked correlation coefficient between the similarity scores obtained using the embeddings and the expert-judged similarity scores.

4.2. Clinical Term Normalization

Table 5 shows the results for the clinical term normalization task on the MCN dataset [27] which was used in the n2c2 2019 shared-task [33]. The first column shows results obtained by the full system. The second column shows the results when the patterns as described in Subsection 3.3.2 were 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.

	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 5. 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 accuracies (%) when the correct answer is 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% [33]. 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 using machine learning methods for the normalization task but it simply used embeddings learned from SNOMED CT to find the most similar concept. It did not even use the training data provided in the MCN corpus other than using it as a source of additional synonyms of clinical terms for exact matching. Yet our system performed competitively and would have secured 7th rank in this shared-task based on the 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 5, one can observe that exact matching alone obtains 76.05% accuracy which is consistent with prior reporting [36]. With exact matching also, sometimes the correct answer is not the first answer but could be the second answer showing that clinical terms are sometimes ambiguous and can exactly match with more than one concept. 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. Although 1% absolute improvement may not look large, it should be pointed out that the margin of improvement is low on this dataset because its post-adjudication inter-annotator agreement was itself low (74.2%). Furthermore, the patterns were not designed to correct every type of error the normalization systems could be making but were designed specifically to correct the errors resulting from the limitations of embeddings that were pointed out in Subsection 3.3.1.

On a single-CPU computer (Intel Core i7-10700T, 2GHz, with 16 GB RAM), the "Exact only" method took 0.01 seconds on average to normalize a clinical term. "Embeddings+Exact" method took 1.31 seconds while "Embeddings+Patterns+Exact" took 1.54 seconds on average to normalize a clinical term. It is not surprising that using embeddings would take much longer computational time than exact matching because it needs to compute cosine similarity of the given clinical term with every clinical term in UMLS to find the top closest ones. However, this step is amenable for parallel processing and could be faster on a GPU-based computer. We note that using patterns only marginally increased the computational time, this is because in our implementation pattern matching is done efficiently by storing the patterns in a hash table.

	Embeddings+Patterns+Exact	Embeddings+Exact	Exact only
Top 1	86.25	86.11	83.45
Top 2	87.23	87.09	84.00
Top 5	87.84	87.84	84.00
Top 10	88.09	88.06	84.00

Table 6. Results on the clinical term normalization subtask of the ShARe/CLEF eHealth 2013 dataset using the embeddings obtained from SNOMED CT using our method. The large number of CUI-less clinical terms were excluded from this evaluation. The numbers are accuracies (%) when the correct answer is within the top 1, 2, 5, and 10 closest concepts according to the system.

Table 6 shows the results on the ShARe/CLEF eHealth 2013 dataset. This dataset has a large number of CUI-less clinical terms (32.7% in the testing dataset) which the exact-matching-only method would trivially answer correctly. We also found that when other methods would normalize such clinical terms, they would sometimes normalize them to their correct CUIs even though the dataset would have CUI-less as the correct answer. Thus other methods would get unfairly penalized. For example, the clinical term "atrioventricular conduction block" is labeled CUI-less in the test set even though its synonym "AV block" is labeled with its correct CUI somewhere else in the test set. Similarly, the clinical terms, "mitral leaflets thickened", "bilateral effusion" and "lv systolic function depressed" are labeled CUI-less in the test set but the same terms are labeled with their correct CUIs in the training set. Hence we decided to only test on the clinical terms with CUIs associated with them (total 3,601) excluding all the CUI-less clinical terms. This setting is not unrealistic because it is testing the ability of a system to normalize clinical terms to their correct CUIs.

The results in Table 6 show a similar trend as the results on the MCN dataset thus showing that the performance of the embeddings learned from SNOMED CT generalizes across datasets. On this dataset, the patterns helped only marginally. Embeddings from ClinicalBERT did slightly better on this dataset obtaining 86.53% accuracy without patterns and 86.67% with patterns. The ShARe CLEF eHealth 2013 dataset was meant for joint named entity extraction and normalization tasks [34], hence most of the past results on this dataset are not comparable because they do not show results separately on the normalization task. However, we did another evaluation in which we also excluded the clinical terms from testing whose concepts are in the training data. In this setting, only the remaining 618 clinical terms were tested. This setting is comparable to the "Unseen concepts" setting from [44] in which a method and embeddings were specifically trained for the normalization task and obtained 71.68% accuracy. Our results are shown in Table 7. The numbers are lower than in Table 6 because this test setting is more difficult given that no concept matches in the training set. For this reason, the exact-matching-only method does much worse. The methods that used embeddings obtained a bigger improvement over exact-matching-only method. There is also a bigger difference from Top 1 to the next top results showing that if the topmost answer is not correct then the correct answer is often near the top. On this setting, embeddings from ClinicalBERT obtained slightly worse results with 69.26% without patterns and 70.06% with patterns. We note that ClinicalBERT embeddings as well as the embeddings obtained using SNOMED CT are general embeddings and were not trained specifically for the normalization task.

	Embeddings+Patterns+Exact	Embeddings+Exact	Exact only
Top 1	70.39	70.06	58.58
Top 2	74.76	73.95	61.17
Top 5	76.54	76.21	61.17
Top 10	77.18	76.86	61.17

Table 7. Results on the clinical term normalization subtask of the ShARe/CLEF eHealth 2013 dataset using the embeddings obtained from SNOMED CT using our method. The large number of CUI-less clinical terms and the clinical terms whose concepts were in training data were excluded from this evaluation. The numbers are accuracies (%) when the correct answer is within the top 1, 2, 5, and 10 closest concepts according to the system.

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 8. 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, none of which is already present in UMLS, the top 5 most similar terms in UMLS found using each type of embeddings are shown.

4.3. Qualitative Comparison

Besides quantitatively evaluating the embeddings obtained using SNOMED CT on two tasks, we qualitatively evaluated them and compared them with corpus-based embeddings. Table 8 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.

5. Limitations and Future Work

The results presented in the previous section show that clinical term embeddings obtained from SNOMED CT using our method capture knowledge about clinical terms well and hence do better than corpus-based embeddings on the clinical term similarity task and competitively on the clinical term normalization task. We expect them to also do well on tasks where sole clinical terms are involved, such as term-based searches and various ontological tasks. However, unlike corpus-based embeddings, they do not capture linguistic knowledge because the method was never trained on text data. Hence these embeddings lack the information about possible surrounding contexts of clinical terms in sentences. As a result, the embeddings obtained from SNOMED CT alone cannot do well on NLP tasks such as named entity recognition which heavily depends on cues from surrounding text to recognize named entities. There is an important lesson here that goodness of embeddings is task-dependent, that is, an embedding that is good for one task may not be good for another task and vice-versa. However, a possible future work will be to suitably combine ontology-based and corpus-based embeddings, including contextual embeddings obtained using recent transformer-based architectures, so that knowledge from both types of sources could be incorporated into embeddings. We expect that a method that fully utilizes both the sources of knowledge will obtain better embeddings.

As was pointed out in the Subsection 3.3.1, both corpus-based and ontological-based embeddings suffer from the limitation that they learn similar embeddings for the terms which have opposite or analogous meanings. In this work, we used a method to learn patterns which captured such terms and then used the patterns to correct possible mistakes caused by embedding-based similarity. There is a lesson here that sometimes a rule-based approach could easily achieve what may be difficult for an embedding-based approach to achieve, hence supplementing embeddings with some rules could be sometimes a viable option. However, alternatively, a more principled approach could be developed that would prevent embedding methods from learning similar embeddings for terms with opposite or analogous meanings. In past, there have been approaches in the general domain to post-process embeddings so that they are dissimilar for antonyms [45]. However, this approach expects a list of antonym terms which is not available for clinical terms. But our learned negative patterns could be treated as such a resource. Hence a possible future work could be to incorporate these patterns in the embedding learning process in order to improve them. We also point out that our learned positive and negative patterns could be potentially useful for other tasks besides clinical term normalization.

6. Conclusions

Traditionally, word embeddings are obtained from text corpora. In this paper, 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 task. 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 a method that automatically learns patterns that indicate whether two clinical terms could mean the same concept or not which were used to supplement the embeddings for the normalization task. However, these embeddings lack linguistic knowledge because they are not trained using text corpora. In future, the two resources of embeddings could be leveraged together to obtain enhanced embeddings.

Acknowledgements

We thank those who created the datasets and the tools used in this work and made them publicly available.

References

1 Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, Soni S, Wang Q, Wei Q, Xiang Y, Zhao B. Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association. 2020 Mar;27(3):457-70.

2 Kalyan KS, Sangeetha S. SECNLP: A survey of embeddings in clinical natural language processing. Journal of biomedical informatics. 2020 Jan 1;101:103323.

3 Schulz C, Juric D. 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 2020 Apr 3 (Vol. 34, No. 05, pp. 8775-8782).

4 Staab S, Studer R, editors. Handbook on ontologies. Springer Science & Business Media; 2010 Mar 14.

5 SNOMED CT. URL: https://www.snomed.org/ Accessed: June 2023.

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

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

8 Choi Y, Chiu CY, Sontag D. Learning low-dimensional representations of medical concepts. AMIA Summits on Translational Science Proceedings. 2016;2016:41.

9 De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P. Medical semantic similarity with a neural language model. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management 2014 Nov 3 (pp. 1819-1822).

10 Baader F, Horrocks I, Sattler U. Description logics. Springer Berlin Heidelberg; 2004.

11 Agarwal K, Eftimov T, Addanki R, Choudhury S, Tamang S, Rallo R. Snomed2Vec: Random Walk and Poincare Embeddings of a Clinical Knowledge Base for Healthcare Analytics. 2019 KDD Workshop on Applied Data Science for Healthcare (DSHealth '19).

12 Chen J, Hu P, Jimenez-Ruiz E, Holter OM, Antonyrajah D, Horrocks I. OWL2vec*: Embedding of owl ontologies. Machine Learning. 2021 Jul;110(7):1813-45.

13 Castell-Díaz J, Miñarro-Giménez JA, Martínez-Costa C. Supporting SNOMED CT postcoordination with knowledge graph embeddings. Journal of Biomedical Informatics. 2023 Mar 1;139:104297.

14 Kate RJ. Automatic full conversion of clinical terms into SNOMED CT concepts. Journal of Biomedical Informatics. 2020 Nov 1;111:103585.

15 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).

16 Faruqui M, Dodge J, Jauhar SK, Dyer C, Hovy E, Smith NA. Retrofitting Word Vectors to Semantic Lexicons. InProceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2015 (pp. 1606-1615).

17 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.

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

19 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.

20 Noh J, Kavuluru R. Improved biomedical word embeddings in the transformer era. Journal of biomedical informatics. 2021 Aug 1;120:103867.

21 Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000 Jul;88(3):265.

22 Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. International Conference on Learning Representations 2013.

23 Goodfellow I, Bengio Y, Courville A. Sequence modeling: recurrent and recursive nets. Chapter 10. Deep learning. MIT Press: 2016.

24 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

25 Dey R, Salem FM. Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international Midwest symposium on circuits and systems (MWSCAS) 2017 Aug 6 (pp. 1597-1600). IEEE.

26 Keras. Francois Chollet and others, 2015, URL: https://keras.io.

27 Luo YF, Sun W, Rumshisky A. MCN: a comprehensive corpus for medical concept normalization. Journal of biomedical informatics. 2019 Apr 1;92:103132.

28 Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, Kingsbury P, Liu H. A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics. 2018 Nov 1;87:12-20.

29 Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics. 2007 Jun 1;40(3):288-99.

30 Hliaoutakis A. Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline. Master's thesis. 2005 Nov 1.

31 Pakhomov SV, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. Journal of biomedical informatics. 2011 Apr 1;44(2):251-65.

32 Pakhomov SV, Finley G, McEwan R, Wang Y, Melton GB. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics. 2016 Dec 1;32(23):3635-44.

33 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.

34 Pradhan S, Elhadad N, South BR, Martinez D, Christensen LM, Vogel A, Suominen H, Chapman WW, Savova GK. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. CLEF (working notes). 2013 Sep 23;1179.

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

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

37 Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA, June 2019.

38 Schulz C, Levy-Kramer J, Van Assel C, Kepes M, Hammerla N. Biomedical Concept Relatedness–A large EHR-based benchmark. In Proceedings of the 28th International Conference on Computational Linguistics 2020 Dec (pp. 6565-6575).

39 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.

40 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.

41 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).

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

43 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.

44 Xu D, Miller T. A simple neural vector space model for medical concept normalization using concept embeddings. Journal of biomedical informatics. 2022 Jun 1;130:104080.

45 Samenko I, Tikhonov A, Yamshchikov IP. Intuitive contrasting map for antonym embeddings. Frontiers in Artificial Intelligence and Applications. 2021;341:502-10.