Concept embedding to measure semantic relatedness for biomedical information ontologies

Junseok Park\textsuperscript{a}, Kwangmin Kim\textsuperscript{a}, Woochang Hwang\textsuperscript{b}, Doheon Lee\textsuperscript{a,\textdagger}

\textsuperscript{a} Department of Bio and Brain Engineering, KAIST, Daejeon, Republic of Korea
\textsuperscript{b} Mlíner Therapeutics Institute University of Cambridge, Cambridge CB2 1TN, UK

\textbf{ARTICLE INFO}

**Keywords:**
UMLS
Similarity
Paragraph vector
Embedding
NLP
Wikipedia

\textbf{ABSTRACT}

There have been many attempts to identify relationships among concepts corresponding to terms from biomedical information ontologies such as the Unified Medical Language System (UMLS). In particular, vector representation of such concepts using information from UMLS definition texts is widely used to measure the relatedness between two biological concepts. However, conventional relatedness measures have a limited range of applicable word coverage, which limits the performance of these models. In this paper, we propose a concept-embedding model of a UMLS semantic relatedness measure to overcome the limitations of earlier models. We obtained context texts of biological concepts that are not defined in UMLS by utilizing Wikipedia as an external knowledgebase. Concept vector representations were then derived from the context texts of the biological concepts. The degree of relatedness between two concepts was defined as the cosine similarity between corresponding concept vectors. As a result, we validated that our method provides higher coverage and better performance than the conventional method.

1. Introduction

Semantic relatedness is a general example of semantic similarity referring to the determination of whether two biological terms are related \cite{1,2}. How semantic relatedness or semantic similarity is calculated is linked to core methods of various technologies, such as bioinformatics, which can distinguish biological terms into meaningful groups, along with the literature-based information retrieval of medical informatics \cite{3,4}. Calculation methods have been applied in various biomedical fields. Boyack et al. \cite{5} clustered numerous biomedical publications according to their similarity using biological terms. Mathur et al. \cite{6} studied disease similarity levels using methods for finding semantic similarity levels between biological processes. Guo et al. \cite{7} used semantic similarity measures to describe direct or indirect interactions within human regulatory pathways. Shah et al. \cite{8} visualized self-organizing maps of biomedical document clusters based on disease concepts.

Semantic relatedness is a measure independent of the hierarchical ontologies of biological terms. It has the advantage of allowing searches of large-scale Metathesaurus and semantic networks developed through the integration of various ontologies \cite{9}, as semantic relatedness is measured by quantifying shared information contents or using context vectors of biological terms, and not depends on a hierarchical ontology and uses “is a” relationships as semantic similarity \cite{10,11}. For instance, the Lesk method calculates the number of common words among definitions of concept pairs \cite{12} while the Vector method generates a first-order co-occurrence matrix for each word and then builds a second-order co-occurrence matrix based on extended definitions of biological terms as a gloss vector \cite{13,14,15}. The gloss vector is used to calculate the semantic relatedness score between biological terms. These methods are indispensable for improving the similarity calculation method used during searches of the Unified Medical Language System (UMLS), which is integrated with various ontologies. However, these methods are known to have lower accuracy than those using hierarchical relationships. Accordingly, the issue of low accuracy would cause no significant differences among vectorized concepts when utilizing them \cite{16}.

The low accuracy occurs because concept definition resources for calculating semantic relatedness from UMLS are inadequate. Insufficient coverage is also known to be a critical obstruction in large-scale biomedical text processing methods \cite{17}. The Vector method attempts to mitigate this problem by extending definition information using the available path information of concepts, but this has had a limited effect; only approximately 6.5% concepts from the 2015AB version of UMLS have definition information in the case of absent path information \cite{9}. Algorithms are another factor. The weakness of the
Lesk method is that it is heavily dependent on dictionary descriptions, though this does not necessarily mean that overlapping must arise when biological terms are semantically related [18]. Although the Vector method is a fine-grained measure that resolves the problems of the Lesk method, it is also bag-of-words approach that considers the sequence information of words in a sentence less [19]. Thus, it essentially has the limitation of low accuracy when determining the semantic relatedness of biological terms.

Here, we propose a concept-embedding model for UMLS semantic relatedness calculations to improve the performance of previous computational semantic relatedness methods based on vectorization with extended data for UMLS. This method consists of generating inexistant UMLS concept definitions as features and creating vectors of concept unique identifiers (CUI) as paragraph vectors. Our method is based on the distributed representations of sentences and documents [20]. We also use a scoring function to determine the relatedness measures through the cosine similarity values between CUI vector pairs generated from our model. Using this preprocessing method and model, we confirm that our approach has better coverage even without path or definition information in UMLS, and we obtain better relatedness measures compared to that by the Vector method [13]. Therefore, our extended definition dataset and semantic relatedness calculation model will contribute to the development of biomedical information retrieval technology in the UMLS Metathesaurus.

Our method is based on the concept of distributed representations of sentences and documents by Le et al. [20]. We also referred to work by McInnes et al. entitled UMLS-Interface and UMLS-Similarity to find limitations and devise the validation process of our method [11].

2. Background

2.1. Unified medical Language System (UMLS)

The Unified Medical Language System (UMLS) was developed as an integrated knowledge resource of terms in the medical field [21]. Various measures pertaining to the relationships among UMLS terms have been implemented in biological studies, such as disease similarity predictions based on biological processes, pharmacovigilance signal detection, and the construction of biomedical question-answer systems [6,22,23]. Identifying the relationships among concepts corresponding to the terms from UMLS is a promising approach by which to understand the relationships among diverse biological concepts. For this purpose, the UMLS Metathesaurus contains information about various biomedical concepts and the relationships between them. The Metathesaurus uses a unique identifier when a concept is added, and it places concepts at the following four levels.

- Concept Unique Identifier (CUI): An identifier for all linked strings (names) from source ontologies that indicate the same meaning (concept)
- Lexical Unique Identifier (LUI): Links for groups of strings that have lexical variants
- String Unique Identifier (SUI): Each unique string that appears in each language in the UMLS Metathesaurus
- Atom Unique Identifier (AUI): Each of the strings in each of the source ontologies

One CUI can contain multiple LUIs, SUIs and AUIs. Each unique AUI is stored in a MRCONSO file (table) that stores concept names and sources based on the UMLS Rich Release Format (RRF). Definitions, which are the attributes of each unique AUI, are stored in a MRDEF file (table). The structure of the unique identifiers and relationships with regard to the RRF file(s) (table(s)) are depicted in Fig. 1.

2.2. Semantic similarity and semantic relatedness

The relationships between UMLS concept pairs fall into two distinct categories according to the definitions of Pedersen et al. [11]: similarity measures and relatedness measures. Similarity measures between a concept pair quantify their closeness in an ontological hierarchy to represent how alike they are. Given that similarity measures are based on path information between biological concepts, they are regarded as ontology-dependent measures [24–35]. While the semantic similarity measure calculates how much two concepts are alike considering an "is a" relationship, semantic relatedness measures between a term pair quantify how the terms are semantically related based on shared information contents between two terms [10]. Because relatedness measures are based on the definition text information of the concepts, they are regarded as ontology-independent measures [12,13].

Relatedness measures have several benefits over similarity measures [9]. One evident benefit is the wider coverage of the relatedness measures. Similarity measures can be calculated only if selected concepts meet the requirements, i.e., path information between the concepts exists. In contrast, even in cases where the concepts have no path information, the relatedness measure can be calculated when the concepts have definition information in UMLS [12]. Moreover, relatedness measures consider the semantic information of the definition texts, which is not considered by similarity measures [13,14].

2.3. UMLS-interface

McInnes et al. developed UMLS-Interface in the form of a Perl package program to provide an API with which to explore locally installed instances of UMLS [36]. The program can be utilized to find information about a CUI, such as its ancestors, depth, definition, extended definition, and all paths to the root and semantic types. The current version of the program is 1.51, and it contains 29 utility programs.

2.4. UMLS-similarity

UMLS-similarity is a Perl package program which provides new similarity/relatedness measures for comparison with existing methods based on UMLS [37]. The program provides similarity/relatedness scores that are computed from UMLS by extracting the concept information and path according to given method and source(s). This program includes an application programming interface (API) and a command line interface (CLI) for users. The current version of the program is 1.47.

2.5. Related works

Semantic similarity and relatedness have been defined in various works, and many methods have been studied accordingly. Rodriguez and Petrakis conducted a study to compute semantic similarity levels from different knowledge sources. They also discussed methods by which to calculate cross-ontology similarity levels using different knowledge sources [38,39].

Pirrò defined similarity and relatedness based on the ontology structure and presented a method to compute their scores. According to Pirrò’s definition, similarity considers only the subsumption relationship between the two concepts (i.e., ‘is-a’), whereas relatedness considers a broader range of relationships (e.g., ‘part-of’) [30].

Banerjee proposed a Lesk measure to determine the relatedness between two concepts. In the Lesk measure, the relatedness between two concepts is determined by the overlap between their gloss definition texts. That is, the relatedness between the two concepts increases as the definition text becomes similar. Fig. 2 shows a simplified example of the calculation of semantic relatedness via the Lesk method. In the Lesk method, the relatedness score is given by the sum of the
squares of the length of the overlap words. In practice, the definitions of the CUI terms themselves as well as the definitions of their related terms are considered. However, because the Lesk method involves simple calculations according to the number of overlapping words, it does not effectively represent the similarity between two definition texts.

To overcome the disadvantages of the aforementioned Lesk measure, Patwardhan proposed the Vector method. In the Vector method, the relatedness between two concepts is calculated as the cosine similarity between the gloss vectors in the word space, which is constructed from the co-occurrence matrix of gloss definition texts [40].

Pointwise mutual information (PMI) is a measure of the association between two features in information theory or statistics. It is conceptually similar to semantic relatedness. While statistical association and semantic relatedness are not equivalent concepts given that a pair of closely related concepts does not necessarily co-occur frequently in the text, previous studies have applied PMI to compute semantic relatedness. Pesaranghader used PMI as an adjunct feature to cut-off co-occurrence data, improving gloss Vector-based relatedness measures [15].
Word2Vec is a natural language processing (NLP) method that vectorizes words by preserving the co-occurrences of frequently appearing words [19,41]. It is based on a distributional hypothesis which states that words appearing in similar locations are similar in meaning [42]. Word2Vec has two methods: the Continuous Bag of Words (CBOW) and the Skip-Gram methods. CBOW predicts a word in the center using the surrounding words, whereas Skip-Gram predicts surrounding words using a center word. In biomedical domain tasks, Skip-Gram reportedly outperforms CBOW [16].

Word2Vec has influence beyond general NLP to calculate semantic similarity scores in the biomedical domain. Yu et al. proposed a method of modifying the context vector representations of medical subject heading (MeSH) terms by using the additional information of UML and the MeSH hierarchy to improve the semantic similarity between terms [43]. Studies of distributed representation showed that Word2Vec is useful given its simplicity and versatility of vector representation when determining medical concept similarity levels and for query expansion and literature-based discoveries in medical informatics [44,45]. In addition, the results of similarity embedded vectors from biomedical datasets without relationship information were comparable to those in previous studies [46].

Glove is also a word-embedding method [47]. Glove preserves the concurrency information of words, similarly to Word2Vec. The inner product of an embedded word vector in Glove is equal to the logarithm of the probability of co-occurrences. In other words, Glove converts the words into vectors by preserving the ratio of the co-occurrence information. Glove improved on the limitations of Word2Vec, which learn co-occurrences in the entire corpus by learning and analyzing only within the window specified by a user. The objective function of Glove is defined as reflecting the statistical information of the entire corpus. However, Word2Vec with a specific language model was observed to perform better than Glove when systematically comparing the similarity and the relatedness of biomedical concepts in the biomedical domain [48].

Doc2Vec is another embedding method for variable fixed-length pieces of text, such as sentences, paragraphs, or entire documents [49]. CBOW and Skip-Gram of Word2Vec correspond to the Paragraph Vector with Distributed Bag of Words (PV-DBOW) and the standard Paragraph Vector with Distributed Memory (PV-DM) concepts. The objective of Doc2Vec is to improve the classification performance between labels by reflecting the characteristics of the input texts. The performance of PV-DM is generally better than that of PV-DBOW, similar to like Word2Vec [49]. The learning method of Doc2Vec consists of receiving a list of words and labels of each sentence and then updating a vector representation of each label of each word set. Yao et al. suggested a method which uses Doc2Vec to obtain the best classification performance for traditional Chinese medical records [50]. The AZTEC platform is an analysis tool for processing multi-omics data and similarity calculations between digital resources using Doc2Vec [51].

To the best of our knowledge, none of the methods discussed above are intended to compute the semantic relatedness between concepts of UMLS. Therefore, in this manuscript, we propose a method which complements concept definitions that are not included in UMLS using an external knowledgebase and calculates the semantic relatedness between the concepts.

3. Methods

We propose a method that solves the coverage limitations of conventional vector-based UMLS relatedness measures and shows improved performance.

We applied two strategies to address the limitations of vector-based UMLS relatedness measures (Fig. 3). First, we extended the definition information of the CUI terms using the Wikipedia database to improve the coverage of the similarity model. Second, we adopted document embedding for vector representations of the CUI terms rather than the bag-of-words approach used by the Vector method to improve the performance of the relatedness measure [19].

3.1. Extension of the definitions of UMLS CUIs for feature vector extraction

The UMLS database contains CUI concept definition information for only a small portion of CUI concepts, limiting the potential coverage of relatedness measures. In the UMLS2015 case, only 162,973 CUI concepts (6.5% of CUI concepts) contain definition information. We used two methods to expand the CUI definition texts not offered by the UMLS database (Fig. 4).

First, we utilized the set of terms related to the given CUI term to obtain text information rather than utilizing only the given term itself. Relationship information was obtained from UMLS. We used the criteria of Liu et al. to distinguish proper related terms; only known relationships between CUI concepts were used [9]. We then selected the hierarchical relationships from UMLS, which consist of parent/child and broader/narrower relationships.

Secondly, we utilized Wikipedia as a source of context texts for the CUI concepts [52]. Wikipedia has been adopted as a source for various vector embedding models, and it has been confirmed as feasible for use with document embedding methods [53,54]. In this study, we used both UMLS definition text and Wikipedia articles to derive CUI embedded vectors.

We used open-source Python APIs to parse data content in Wikipedia, and pages from Wikipedia were parsed during August of 2017. Articles from Wikipedia were extracted with the following priority: (1) if there is a Wikipedia article which has a title that exactly matches a CUI term, the corresponding Wikipedia article was extracted. (2) If there is a redirection link that matches a CUI term, the redirected article was extracted. (3) If there is any search suggestion for a CUI term, even in the absence of a precisely matching article, the first suggestion was extracted (Fig. 5).

The Wikipedia article text has features which differ from those of the CUI definition text, making it inappropriate to combine the two texts directly in the expanded definition text. First, Wikipedia articles are much larger on average than CUI concept definition text, and they are more diverse in terms of size as well. Second, the ranges of information contained in the articles are not uniform. Articles on well-known and general subject matter tend to be described from more diverse perspectives, while articles on uncommon subject matter tend to deal only with brief information. For example, the Wikipedia entry for 'Coronary artery disease' consists of eleven sections, including causes, pathophysiology, diagnosis, screening, and prevention of the disease. In contrast, the Wikipedia article on thyrotoxicosis factitia, a type of hyperthyroidism, consists of only two paragraphs.

To reflect the Wikipedia and CUI concept definition information uniformly, we collected only lead paragraph text instead of the full text of each Wikipedia article. Although Wikipedia documents are created by users’ free participation, the general structure of each document follows a certain format, which is also specified in the Wikipedia user guidelines [37]. The lead paragraph of a typical Wikipedia article presents the definition and gives a brief introduction of the topic [55]. Because lead paragraphs of the articles contain information similar in terms of quality to CUI concept definitions and are uniform in size, we considered lead paragraphs to be appropriate to combine the two texts into an expanded definition text. Fig. 6 shows the examples of UMLS definition texts and the lead paragraphs extracted from Wikipedia for certain CUI terms. Even in the case of uncommon terms such as 'Thyrotoxicosis factitia', in which the CUI definition text does not exist, lead paragraphs can be extracted from Wikipedia.

Utilization of the text information from Wikipedia improved our dataset in terms of both quantity and quality. We extracted 946,785 extended definition texts from Wikipedia with this procedure. Finally, we obtained the text information of CUI terms by concatenating the texts obtained from the two methods above, the title, the UMLS
definition, and the lead paragraph of the Wikipedia article from the set of terms related to the given CUI term.

3.2. Model generation with a document embedding method for UMLS CUI feature vector generation

Word embedding, a method of representing a word with a vector, is done on the assumption of what is termed a distributional hypothesis [42]. General and basic word-embedding methods derive from the bag of words concept. These methods, however, cannot reflect the semantic information of words due to the absence of word order information. This represents a limitation that hinders our understanding of the differences or similarities among words constituting sentences [56]. Therefore, many researchers have devised word-embedding methods which represent the meaning of the word itself in a reduced multi-dimensional space. In chronological order, these methods are the Neural Net Language Model (NNLM), the Recurrent Neural Network based Language Model (RNNLM), and the architecture of Continuous Bag-of-Words (CBOW) and Skip-Gram [19,57,58] methods. A document embedding method capable of expressing semantic similarity based on the above methods was also developed using dense vectors of variable-length sentences, paragraphs, and documents [20].

We analyzed the relationship between a CUI and a definition of the CUI and concluded that it is a label relationship, not an inclusive relationship. We found that the CUI is a label and that the definition of a CUI is a feature. We then utilized a previous method based on state-of-the-art word and document embedding techniques.

We used a UMLS definition as a paragraph. All words in a paragraph were processed by a continuous word-embedding method with the Skip-Gram architecture. We built a model for obtaining a paragraph vector from embedded words in a paragraph by PV-DBOW. The CUI was used as the paragraph ID in the model (Fig. 7).

As a model development framework, we used DeepLearning4J 0.9.1 and modified it to achieve optimal performance for solving the problem at hand [59]. Co-occurrence information between the words in the definitions is preserved by embedding the words using Word2Vec before the model is trained by PV-DBOW [19]. In order to compensate for the differences between UMLS and Wikipedia, we used AdaGrad to apply different gradients among the features [60]. It works as a normalization function by decreasing the effective learning rate of the weights with a high gradient value or by increasing the effective learning rate of slightly updated weights or weights with a low gradient value. We defined the number of words to be processed at one time in the definitions of the CUIs as 1000 and reflected the number in the batch size, with a window size of 5. The word count in the corpus was set to 1 to process all words in the definitions. Subsequently, we optimized the hyperparameters of the learning rate, layer size and epoch. The learning rate is the step size for each update, the layer size is the number of dimensions in the vector space, and the epoch refers to all of the learning samples that have been used once [61]. There were ten iterations for batch updating of the data. We set the number of epochs to 1 and adjusted the single hidden layer size from 100 to 500 within 100 steps. We also used a range of different learning rates from 0.001 to 0.03 with a 0.001 step size, and we selected a layer size of 300 and a learning rate of 0.025 to minimize the loss function of PV-DBOW [37,60,62]. We used 3, 5, 10, 20 and 30 epochs. Finally, we compared the resulting coefficients with the benchmark set with 30 medical term pairs by means of the Spearman’s rank correlations [63] (Table 1).

A paragraph vector is obtained using a CUI input into the generated model. A relatedness score is calculated using the cosine similarity calculation method with Eq. (1). According to the definition of cosine similarity, the range of this score is −1 to 1, where −1 indicates that the two terms are opposite, 0 means that they are independent, and 1 means that they are equal to each other.

\[
\text{Relatedness score } (CUI_i, CUI_j) = \frac{CUI_i \cdot CUI_j}{||CUI_i|| ||CUI_j||}
\]

4. Results

We conducted experiments to assess the coverage expansion, the accuracy improvements and result significance levels of the proposed method. In order to compare the coverage of our method with that of
the Vector method, classified as the relatedness measure, we randomly generated CUI pairs and compared a number of existing relatedness scores. We demonstrate the performance of our method by comparing vector measures using Spearman’s rank correlation with manually ranked CUI pairs. Our method has 4.77% more coverage compared to the Vector method. We also found relatively higher performance by the proposed method as compared to the Lesk and Vector methods based on the benchmark set. The result process was derived from Liu’s proposed process [9]. Finally, we validated the result significance levels of our model by comparing the benchmark set with a random set. CUIs with path information in UMLS2015AB, UMLS-Similarity 1.45 and UMLS-Interface 1.47 were applied to measure the performance of our method. The detailed results are discussed below.

4.1. Coverage comparison of randomly selected CUI pairs

In this experiment, we attempted to identify the coverage when calculating a relatedness score that had not been calculated before using the expanded definition from Wikipedia. The vector method calculates the relatedness score by aggregating the definition information of a CUI and extending the definition information using the available path information from the CUI. We created a total of 998,543 definitions. By extending the definition information with Wikipedia, we obtained 964,785 new definitions after combining 162,973 existing definitions and accounting for overlap. For a coverage difference comparison, we generated ten randomly selected sets of 1000 CUI pairs.

![Fig. 6. Case examples of UMLS definition texts and lead paragraphs extracted from Wikipedia.](image)

![Fig. 7. An example of the CUI embedding architecture of the model.](image)

| Table 1 |
|---|---|---|---|---|---|
| Score/epoch size | 3 | 5 | 10 | 20 | 30 |
| Spearman’s rank correlation coefficients with the benchmark set from Mayo Clinic | 0.745 | 0.777 | 0.795 | 0.725 | 0.676 |
We summed number of scores of 0, indicating no definition information in the vector method, and NaN, indicating definition information identical to the vector cases in our method, according to each group. We then calculated the average number of results for all trials (Table 2). The results show that we have approximately 4.77% more coverage than that by the Vector method.

The results demonstrate the importance of complementing the incomplete UMLS, with improved features when comparing each CUI definition [64,65]. In particular, we used Wikipedia for features, meaning that each definition is potentially reflected by various authors. The potential to reflect its characteristics when embedding the CUI is assumed. This assumption would also include the possibility of more clearly calculating the semantic relatedness of non-related CUIs. Thus, the results here can serve as an important resource in the biomedical text mining field.

### 4.2. Performance comparison with previous works

Semantic relatedness does not simply measure the “is-a” relationship between two biological terms, instead aiming mainly to measure contexts and meanings. Therefore, the development of a semantic-centered comparison method similar to the human thinking process is necessary [36]. To achieve this purpose, we validated that our model shows results similar to those of biomedical benchmark data created by human experts. Benchmark data which contains the semantic relatedness between two biological terms cannot be defined using a generic score and instead uses a rank that is estimated to be relatively close to other term pairs [66]. With regard to measuring semantic relatedness, this process cannot be done by comparing the proposed method with the raw scores of term pairs evaluated by human experts because with semantic relatedness, the two scores tend to change simultaneously, but not in the same way at the same rate compared to the proposed method [67]. Therefore, we used Spearman’s correlation coefficient, which is based on ranked values for each score, to evaluate the relationship between ranks, which represent a sequential variable. We compared the performance of our method with Spearman’s rank correlation coefficients among our model, previous methods, and a benchmark set to confirm the performance improvement.

First, we used a CUI pair list from a dataset by Pedersen et al. as a benchmark set [63]. This dataset consists of 30 medical term pairs. Each pair was manually evaluated by nine medical coders and three physicians at the Mayo Clinic. Evaluation scores are from (4.0), practically synonymous, to (1.0), unrelated on a 1.0 scale. We used the Spearman’s rank correlation coefficients among Liu’s relatedness measure results as basis ranks to test the performance of the proposed method [9]. We compared the coefficient results from the Lesk and Vector methods with those from our method (Table 3). Table 3 contains the ranks from each method. Our method shows significantly higher

| Term of CUI1 | CUI1 | Term of CUI2 | CUI2 | Rank | Lesk | Vector | Concept embedding |
|-------------|------|-------------|------|------|-----|--------|------------------|
| Renal failure | C0035078 | Kidney failure | C0035078 | 1 | 4 | 1 | 1 |
| Heart | C0018787 | Myocardium | C0027061 | 2 | 11 | 5 | 3 |
| Stroke | C0038454 | Infarct | C0021308 | 3 | 14 | 11 | 6 |
| Abortion | C0156543 | Miscarriage | C0000786 | 4 | 13 | 14 | 8 |
| Deletion | C0011253 | Schizophrenia | C0036341 | 5 | 8 | 7 | 4 |
| Congestive heart failure | C0018802 | Pulmonary edema | C0034063 | 6 | 12 | 10 | 7 |
| Metastasis | C0027627 | Adenocarcinoma | C001418 | 7 | 15 | 13 | 14 |
| Calcification | C0175895 | Psoriasis | C009814 | 8 | 29 | 29 | 20 |
| Diarrhea | C0011991 | Stomach cramps | C0344375 | 9 | 19 | 22 | 11 |
| Mitral stenosis | C0026269 | Atrial fibrillation | C0004238 | 10 | 17 | 8 | 5 |
| Chronic obstructive pulmonary disease | C0024117 | Lung infiltrates | C0024109 | 11 | 6 | 16 | 10 |
| Rheumatoid arthritis | C0003873 | Lupus | C040974 | 12 | 3 | 2 | 2 |
| Brain tumor | C0006118 | Intracranial hemorrhage | C0151699 | 13 | 20 | 25 | 16 |
| Carpel tunnel syndrome | C0007286 | Osteoarthritis | C0029408 | 14 | 1 | 4 | 15 |
| Diabetes mellitus | C0011849 | Hypertension | C0020538 | 15 | 2 | 12 | 13 |
| Acne | C00702166 | Syringe | C0039142 | 16 | 26 | 19 | 19 |
| Antibiotic | C0003252 | Allergy | C0020517 | 17 | 7 | 9 | 22 |
| Corticosteroid | C0010137 | Total knee replacement | C0086511 | 18 | 25 | 27 | 28 |
| Pulmonary embolus | C0034065 | Myocardial infarction | C0027051 | 19 | 9 | 6 | 17 |
| Pulmonary fibrosis | C0034069 | Lung cancer | C0242379 | 20 | 5 | 17 | 9 |
| Cholangiocarcinoma | C006698 | Colonoscopy | C0000978 | 21 | 18 | 18 | 12 |
| Lymphoid hyperplasia | C033997 | Lymphoid cancer | C0007107 | 22 | 27 | 23 | 27 |
| Multiple sclerosis | C0003795 | Psychosis | C003975 | 23 | 10 | 15 | 18 |
| Appendicitis | C0003615 | Osteoporosis | C0029456 | 24 | 16 | 13 | 30 |
| Rectal polyp | C0034877 | Aorta | C003483 | 25 | 30 | 30 | 25 |
| Xerostomia | C0045352 | Alcoholic cirrhosis | C0028991 | 26 | 22 | 21 | 26 |
| Peptic ulcer disease | C0039920 | Myopia | C002792 | 27 | 23 | 20 | 29 |
| Depression | C0011581 | Cellulites | C0007642 | 28 | 21 | 24 | 21 |
| Varicose vein | C0042345 | Entire knee meniscus | C0224701 | 29 | 28 | 28 | 24 |
| Hyperlipidemia | C0020473 | Metastasis | C0027627 | 30 | 24 | 26 | 23 |

| Spearman’s rank correlation coefficients | 1 | 0.450 | 0.586 | 0.795 |

We compared the performance of our method with Spearman’s rank correlation coefficients among our model, previous methods, and a benchmark set to confirm the performance improvement.
correlation coefficients with the rank of the benchmark set compared to the Lesk or the Vector method, indicating that our method is capable of higher performance than previous relatedness measure methods.

The second benchmark set consists of 36 biomedical term pairs extracted from similarity results with eight medical experts who used evaluation scores ranging from 0 (non-similar) to 1 (synonyms) [36, 68]. We mapped the biomedical terms to the CUIs using Metamap, which maps input biomedical terms to a CUI based on the UMLS2015AB database [69]. When a term maps to more than one CUI, we selected a CUI which belongs to the Disease or Syndrome (T047) category in terms of the UMLS semantic type. For example, adenovirus is mapped to C0001483 (T005:Virus), C0001486 (T047:Disease or Syndrome).

Table 4
Comparison of relatedness scores from the lesk, vector and concept-embedding methods for 36 biomedical term pairs.

| The benchmark set from 8 experts | Lesk | Vector | Concept embedding |
|----------------------------------|------|--------|-------------------|
| Term of CUI1 | CUI1 | Term of CUI2 | CUI2 | Rank | Rank | Rank | Rank |
| Chicken Fox | O0008049 | Varicella | C0042338 | 1 | 15 | 20 | 11 |
| Antibiotics | O003232 | Antibacterial Agents | C0279516 | 2 | 12 | 17 | 13 |
| Measles | O0025007 | Rubella | C0025007 | 3 | 1 | 1 | 1 |
| Pain | O0030193 | Ache | C0234238 | 4 | 26 | 29 | 29 |
| Malnutrition | O162429 | Nutritional Deficiency | C0162429 | 4 | 6 | 1 | 1 |
| Down Syndrome | O013080 | Trisomy 21 | C0557167 | 4 | 23 | 25 | 18 |
| Breast Feeding | O006147 | Lactation | C0022925 | 7 | 18 | 26 | 21 |
| Seizures | O006572 | Convulsions | C006572 | 7 | 3 | 1 | 1 |
| Myocardial Ischemia | O151744 | Myocardial Infarction | C0027051 | 9 | 11 | 9 | 16 |
| Carcinoma | O007099 | Neoplasm | C0027651 | 9 | 13 | 10 | 30 |
| Migraine | O149931 | Headache | C0018681 | 11 | 4 | 11 | 12 |
| Urinary Tract Infection | O002029 | Pyelonephritis | C0034186 | 12 | 17 | 16 | 6 |
| Failure to Thrive | O0015544 | Malnutrition | C0162429 | 13 | 20 | 21 | 19 |
| Vaccines | O0042210 | Immunity | C0020964 | 14 | 30 | 18 | 15 |
| Psychology | O0039099 | Cognitive Science | C0375442 | 14 | 28 | 13 | 7 |
| Hepatitis B | O0019613 | Hepatitis C | C0208407 | 16 | 2 | 4 | 5 |
| Pulmonary Stenosis | O0034089 | Aortic Stenosis | C0003507 | 17 | 19 | 7 | 8 |
| Hypertension | O0020538 | Kidney Failure | C0035078 | 18 | 16 | 19 | 17 |
| Diabetic Nephropathy | O001881 | Diabetes Mellitus | C0011849 | 18 | 9 | 8 | 14 |
| Lactose Intolerance | O0022951 | Irritable Bowel Syndrome | C0022104 | 20 | 5 | 6 | 10 |
| Anemia | O002871 | Deficiency Anemia | C0041782 | 21 | 27 | 22 | 4 |
| Adenovirus | O001486 | Rotavirus | C0035870 | 21 | 36 | 28 | 20 |
| Hypothyroidism | O0020676 | Hyperthyroidism | C0020550 | 23 | 8 | 5 | 9 |
| Sarcoïdosis | O0036202 | Tuberculosis | C0041296 | 23 | 10 | 12 | 25 |
| Asthma | O004096 | Pneumonia | C0032285 | 25 | 7 | 15 | 23 |
| Neonatal Jaundice | O0022353 | Sepsis | C0243026 | 26 | 25 | 24 | 22 |
| Otitis Media | O0020882 | Infantile Colic | C0266836 | 27 | 34 | 36 | 26 |
| Bacterial Pneumonia | O004626 | Malaria | C0024930 | 27 | 14 | 34 | 24 |
| Osteoporosis | O0029456 | Patent Ductus Arteriosus | C349549 | 27 | 35 | 30 | 32 |
| Hyperlipidemia | O0020473 | Hyperkalemia | C0020461 | 27 | 32 | 14 | 27 |
| Amino Acid Sequence | O002518 | Antibacterial Agents | C0279516 | 31 | 33 | 35 | 36 |
| Dementia | O001265 | Atopic Dermatitis | C0016115 | 32 | 31 | 31 | 33 |
| Syndrome | O009082 | Congenital Heart Defects | C0152021 | 32 | 22 | 32 | 35 |
| Anemia | O002871 | Appendicitis | C0030165 | 34 | 24 | 23 | 28 |
| Meningitis | O0025289 | Tricuspid Atresia | C0243002 | 34 | 29 | 33 | 31 |
| Simavitin | O0037199 | Mental Retardation | C0025362 | 34 | 21 | 27 | 34 |

Spearman’s rank correlation coefficients

| | Benchmark set | Random sets |
|-----------------|---------------|--------------|
| Spearman’s rank correlation coefficients | 1 | 0.467 | 0.510 | 0.633 |

Fig. 8. Results of a comparison between the benchmark set and random sets in a box plot.
Syndrome), and C1552907 (T129: Immunologic Factor, T116: Amino Acid, Peptide, or Protein, T121: Pharmacologic Substance), and we chose C0001486. Depending on whether the semantic types of the CUIs are duplicated or not in T047, we used a CUI which has a name most similar to that of the concept of a term. For example, antibiotics are mapped to C0003232 (Antibiotics), C0003237 (Antibiotics, Antituberular) and C3540704 (Antibiotics for systemic use), and all CUIs belong to the semantic type of antibiotics (T195). In this case, we selected C0003232, which has a name most similar to the term (Antibiotics). In addition, if a pair of CUIs is duplicated and one of the terms is mapped to more than one CUI, we selected another CUI of the term. For example, Down’s syndrome is mapped to C00013080 and Trisomy is mapped to more than one CUI, we selected another CUI of the term. In this case, Trisomy 21 is mapped to C001380 and C3537167. Table 4 summarizes the results from the Lesk, Vector and concept-embedding methods after the generation of the CUI pairs. Table 4 also indicates that the concept-embedding method has better performance, as shown in Table 3.

Our results show that applying state-of-art technologies in the field of deep learning can improve performance outcomes. We used PVBOW, rooted in the Skip-Gram technique of Word2Vec, to develop the model. We were able to improve the performance with improved Skip-Gram algorithms, unlike the Lesk and Vector methods. For instance, multi-prototype Skip-Gram, which maps identical words to different vectors if they are used with different meanings, or adaptive Skip-Gram, which reflects semantic differences from the same words depending on the relationships between other words in a sufficient number of processed corporuses, can be utilized to improve the performance [70,71].

4.3. Similarity score significance validation

We conducted this experiment to verify the significance of the results of the proposed method. We compared the score distributions from the model between a benchmark set and random sets. The score distributions illustrate that the sets are highly distinguishable, which implies that the model is statistically significant. We verified our method based on Wang et al., and they validated their method with a benchmark set and with random sets [72]. The benchmark set consisted of 70 highly similar pairs manually curated from 47 diseases [6, 73, 74], similar to their approach, we validated our model by calculating relatedness scores on a benchmark set and another 7000 random sets. We generated a random set with 70 pairs for 1000 iterations from the MRCUI table of UMLS2015AB. We assumed that if the pairs have high similarity, high relatedness scores would also be assigned on average, while if the pairs are randomly selected, scores lower than those of pairs with high similarity would be assigned. We examined average relatedness scores from both sets on our model. As a result, the average relatedness score of the benchmark set was 0.205 and that of the random sets was 0.021. In Fig. 8, the differences between the benchmark set and the random sets are clear. This confirms that our model generates significance scores if the pairs have some degree of similarity between the pairs.

5. Conclusion

This study proposes a concept-embedding model for UMLS semantic relatedness calculations that uses UMLS concept definitions as features. The main contribution of this research lies in how the proposed method calculates reliable semantic relatedness between UMLS concept pairs regardless of whether the concepts have path information. Moreover, compared to existing context-based relatedness measures, we obtained improved coverage by collecting more extensive context texts. Compared to state-of-the-art methods, our method produces more extensive coverage and shows better performance outcomes on UMLS sets. We also adopted Wikipedia as a knowledge base to extend the context text. In the future, it would be meaningful to utilize other extended corpora to obtain rich text sources. Furthermore, it is notable that our semantic relatedness model can potentially be implemented for the biomedical information retrieval of UMLS Metathesaurus terms.

In order to validate our method, we compared our coverage and performance results with those from previous studies. We demonstrated that we can resolve the limited coverage problem unaddressed by the Vector method. We also found that we can obtain better results by comparing Spearman’s rank correlations between the scores by previous methods and that by our model. In the results, we show that the coverage improved by 4.77% on average through a random CUI pair generation test, and we prove the superior performance of our model compared to the correlation coefficients of existing relatedness measure methods.

In conclusion, the proposed UMLS semantic relatedness calculation method is a promising method for finding relationships between UMLS concept pairs. Moreover, while we have focused on UMLS similarity in this study, we suggest that our method can also be applied to calculate the degrees of similarity between other biomedical terms.

Declaration of interests

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

Acknowledgements

This work was supported by the Bio-Synergy Research Project (NRF-2012M3A9C4048758) of the Ministry of Science, ICT and Future Planning through the National Research Foundation of Korea. We are also grateful to Shin Kim for the many discussions about this research.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103182.

References

[1] M. Ben Aouicha, M.A. Hadj Taieb, A. Ben Hamadou, SISR: System for integrating semantic relatedness and similarity measures, Soft Comput. 22 (2018) 1855–1879.
[2] Y. Liu, B.T. McInnes, T. Pedersen, G. Melton-Meaux, S. Pakhomov, Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet, in: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, ACM, Miami, Florida, USA, 2012, pp. 363–372.
[3] L. De Vine, G. Zuccon, B. Koopman, L. Sibton, P. Bruna, Medical semantic similarity with a neural language model, in: Proceedings of the 23rd ACM international conference on conference on conference on information and knowledge management, ACM, 2014. pp. 1819–1822.
[4] P. Agarwal, D.B. Sears, Can literature analysis identify innovation drivers in drug discovery? Nature Reviews Drug Discov. 8 (2009) 865.
[5] K.W. Boyack, D. Newman, R.J. Dubon, R. Klavans, M. Patel, J.R. Biberstine, et al., Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches, PLOS One 6 (2011).
[6] S. Mathur, D. Dinakarpandian, Finding disease similarity based on implicit semantic similarity, J. Biomed. Inform. 45 (2012) 363–371.
[7] X. Guo, R. Liu, C.D. Shriver, H. Hu, M.N. Lieberman, Assessing semantic similarity measures for the characterization of human regulatory pathways, Bioinformatics 22 (2006) 967–973.
[8] S. Shah, X. Luo, Exploring diseases based biomedical document clustering and visualization using self-organizing maps, in: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom): IEEE, 2017, pp. 1–6.
[9] Y. Liu, B.T. McInnes, T. Pedersen, G. Melton-Meaux, S. Pakhomov, Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet, in: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, ACM, 2012, pp. 363–372.
[10] M.A. Hadj Taieb, M. Ben Aouicha, A. Ben Hamadou, A new semantic relatedness measurement using WordNet features, Knowledge Inform. Syst. 41 (2014) 467–497.
[11] B.T. McInnes, T. Pedersen, S.V. Pakhomov, UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity, AMIA Ann. Symp. Proc.: Am. Med. Inform. Assoc. (2009) 431.
[12] S. Banerjee, T. Pedersen, Extended gloss overlaps as a measure of semantic relatedness, Jcenci (2003) 865–810.
