Applied Informatics

RESEARCH Open Access

Concept embedding-based weighting scheme for biomedical text clustering and visualization

Xiao Luo[1*] and Setu Shah[2]

*Correspondence:
luo25@iupui.edu
[1] Department of Computer Information Technology, IUPUI, Indianapolis, USA
Full list of author information is available at the end of the article

Abstract

Biomedical text clustering is a text mining technique used to provide better document search, browsing, and retrieval in biomedical and clinical text collections. In this research, the document representation based on the concept embedding along with the proposed weighting scheme is explored. The concept embedding is learned through the neural networks to capture the associations between the concepts. The proposed weighting scheme makes use of the concept associations to build document vectors for clustering. We evaluate two types of concept embedding and new weighting scheme for text clustering and visualization on two different biomedical text collections. The returned results demonstrate that the concept embedding along with the new weighting scheme performs better than the baseline tf–idf for clustering and visualization. Based on the internal clustering evaluation metric-Davies–Bouldin index and the visualization, the concept embedding generated from aggregated word embedding can form well-separated clusters, whereas the intact concept embedding can better identify more clusters of specific diseases and gain better F-measure.

Keywords: Neural networks, Concept embedding, Biomedical text clustering and Visualization

Introduction

Active research and practice in the medical domain has generated pervasive text files, articles, and documents, which include MEDLINE—the largest biomedical text database, clinical notes in the Electronic Health Records, descriptions of clinical trials, and so on. In order to efficiently discover, search, and access the knowledge within all these text content, there is a continuous need for developing innovative techniques and algorithms for text representation, clustering, and visualization.

Within the biomedical and clinical text files, one medical concept might be represented in different forms or in abbreviations. For example, 'Diabetes Mellitus Type 2' could be represented as 'DM2' or 'Type II Diabetes' in different text files. This happens often in the clinical notes within the Electronic Health Records (EHR), because clinicians have their own preferences of recording notes. On the other hand, some medical concepts might be highly correlated. For example, 'Hypertension' often co-occurs with 'Stroke.' Hence, the co-occurrences and semantic similarities between

words and phrases are important for capturing the similarity of content within the context of biomedical document clustering. In order to capture the semantic similarities between words and phrases, previous research on text representation uses existing ontology such as MeSH or WordNet to identify the semantic relationships (Logeswari and Premalatha 2013; Yoo et al. 2006; Zhang et al. 2007). However, the ontologies present the entity and attribute relationships in a hierarchy without emphasizing the co-occurrences of the concepts. On the other hand, the ontologies often do not include the different representations of the same concept. In recent years, the distributed representation which is also called word embedding gained interests in the research areas of text mining, natural language processing, and health informatics (Mikolov et al. 2013; Moen and Ananiadou 2013; Tulkens and Daelemans 2016). The word embedding emphasizes the co-occurrences of the words based on a given text collection. There are different ways to generate the distributed vector representations, which include probabilistic models (Globerson et al. 2007) and dimensionality reduction on the word co-occurrence matrix (Levy and Goldberg 2014). The neural network is the new and most recent technique to generate the word embedding. It has been recently studied for biomedical text classification and clustering (Tulkens and Daelemans 2016; Kim and Cho 2017; Zhu et al. 2017), where word is the basic unit for the text documents and the word embedding is learned through neural networks. However, in the biomedical domain, clinical or medical concepts often contain more than one word. Hence, it is necessary to investigate the concept embedding that can be learned through the neural networks. We hypothesize that if the text documents are well represented by making use of concept embedding, it can improve biomedical text clustering and visualization.

In this research, we propose and evaluate a framework for biomedical text clustering and visualization based on the concept embedding of diseases. This proposed framework contains four major components: (1) extracting the phrases of diseases; (2) generating the concept embedding through neural networks; (3) constructing text document representations based on a new weighting scheme; and (4) text document clustering and visualization. Our major contributions include evaluation of concept embedding learned through the neural networks and a new weight scheme based on the concept embedding for biomedical text document representation, clustering, and visualization.

The k-means (Hartigan and Wong 1979) algorithm is applied to cluster the newly represented text documents. To visualize the distribution of the clusters on the two-dimensional space, t Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008) is employed after applying the Principal Component Analysis (Pearson 2008) to reduce the original dimensions of the document vectors. To evaluate the quality of the clusters, the internal evaluation metric-Davies–Bouldin index (DBindex) (Davies and Bouldin 1979) and external evaluation metric-F-measure (Van Rijsbergen 1979) are used. Two biomedical text collections, Trecgen and PubMed Open Access, have been used for evaluation. The results demonstrate that the weighting scheme based on the generated concept embedding works much better than baseline tf–idf for the task of biomedical text clustering and visualization.

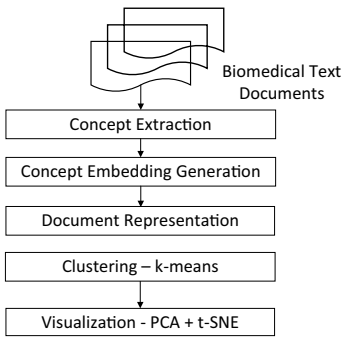

Fig. 1 Overview of the proposed framework

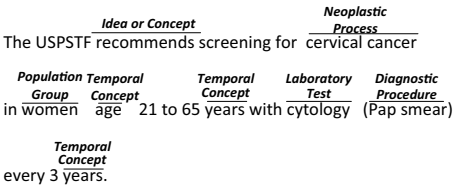

Fig. 2 Semantic mapping of using UMLS MetaMap

Overview of the system framework

Usually, a biomedical text document contains concepts of diseases, medications, treatments, symptoms, and so on. In this project, we focus on evaluating the concept embedding for diseases, given that biomedical literature searches often relate to certain diseases.

In this research, the concept extraction component identifies and extracts the concepts of diseases through using the Unified Medical Language System (UMLS) MetaMap (Shah and Luo 2018). Then, the concept embedding is generated by feeding the extracted concepts through the neural networks. We observe that the larger the text collections, the semantic and syntactic associations of the different concepts can be more accurately captured by the concept embedding. So, the input to the neural networks contains more than one biomedical text collection. After generating the concept embedding, a new weighting scheme is proposed to construct the vector representation of the documents. The new weighting scheme is based on the association scores that are calculated by measuring the distances between the concepts using the embedding. The k-means algorithm is used to cluster the vectors of the documents. Finally, the distributions of the clusters are visualized through the *t*-SNE technique. Figure 1 demonstrates the overall process of the proposed framework. The followed sub-sections detail the methodologies and techniques involved.

Concepts extraction

UMLS MetaMap (Shah and Luo 2018) is a natural language processing tool that uses various sources like UMLS Metathesaurus (Fact Sheet and SNOMED CT) to map the phrases or terms in the text to different semantic types. Figure 2 provides an example of MetaMap's mapping of text into semantic types. In this research, if a term or phrase

has been mapped to semantic types 'Disease or Syndrome' or 'Neoplastic Process,' the corresponding phrase in the lexicon produced by MetaMap is extracted as a concept of disease.

Generation of concept embedding

Generating word embedding through using neural networks (Mikolov et al. 2013) has drawn attention in the areas of natural language processing and machine learning (Tulkens and Daelemans 2016) (Kim and Cho 2017). There are two models of the neural networks for generating the word embedding (Mikolov et al. 2013). One is Continuous Bag of words (CBOW) model, the other is Skip-Gram model. In this research, the Skip-Gram model is employed. The neural network architecture of the Skip-Gram is provided in Fig. 3. It is a standard three-layer neural network.

For word embedding, the input to the neural network are words that are represented as vectors. The length of the vector is the number of words in a text document collection. The element in the vector corresponding to the input word is set to 1, the rest elements are set to 0. The number of neurons at the hidden layer can be defined and tuned depending on the application. The number of output vectors to be evaluated relies on the defined number of context words of an input word. The context words are the words that occur within a specific sliding window of the input word in a sentence. For example, if the window size is 1, there will be two context words in total. One context word occurs before the input word, and the other occurs after the input word in a given sentence. The context words are also represented in the same vector format as the input words. The output vectors will be

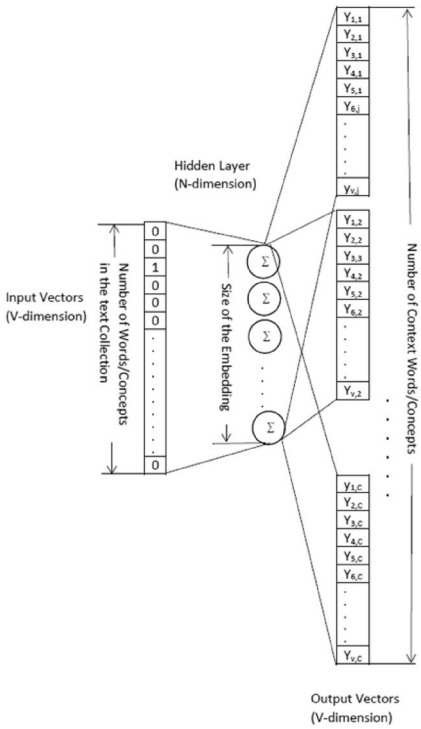

Fig. 3 Skip-Gram model

evaluated against the context words after applying a softmax layer which is used to sum the probabilities obtained in the output vectors to 1. After applying the softmax layer to the output vectors, the differences between the vectors of the context words and the output vectors will be calculated as a summed error vector and propagate back to update the weights of the neural network. After the training process, the weights between the input and the hidden layer are taken as the word embedding. The word embedding preserves the distances between words so that the words that have semantic and syntactic associations in the raw text collection are close to one another.

In this research, we use this Skip-Gram model to generate the concept embedding. Two approaches have been explored: (1) generate the concept embedding by aggregating the word embedding, (2) generate the intact concept embedding, which means the input and output of the Skip-Gram model are vectors for the extract concepts.

Aggregated word embedding

If a concept consists of more than one word, the concept is represented as an element-wise sum of the distributed vectors presenting all words within the concept. For example, the concept 'lung cancer' is calculated as the element-wise sum of vectors of 'lung' and 'cancer.' Given a concept that consists of n words, each word is represented as vector $WordV$, and the vector for the concept is shown as Eq. (1).

$$ConceptV = \sum_{k=1}^{n} WordV_k. \tag{1}$$

Intact concept embedding

The intact concept embedding are generated by considering each concept as a single entity. The structure of the neural network is the same, the concepts are the input to the neural network and the output vectors are corresponding context concepts. Thus, the concept embedding $ConceptV$ can be directly generated.

Concept association measure

Since each concept is represented by a vector $ConceptV$, the association scores between the concepts can be calculated through a distance measure. If the association scores are stored in a matrix S, given a total of M concepts extracted from the raw biomedical text collection, each entry $s_{i,j}$ in the matrix S represents the association score between concept $ConceptV_i$ and $ConceptV_j$. In this research, the Cosine distance (Eq. 3) is used to calculate the association scores.

$$S_{L,L} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,L} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ s_{L,1} & s_{L,2} & \cdots & s_{L,L} \end{pmatrix} \tag{2}$$

$$S_{i,j} = \frac{ConceptV_i \cdot ConceptV_j}{||ConceptV_i||_2 ||ConceptV_j||_2} \tag{3}$$

Based on our previous research (Shah and Luo 2018), both concept embedding generation approaches can capture the associations between concepts. Table 1 provides examples of disease concepts and the their top 3 closest concepts by using Trecgen data set (described in "Data set" section) as the training text collection. The association scores based on the intact concept embedding approach are lower than those based on the aggregated word embedding approach. The reason could be that the intact concept embedding measures the semantic similarities of the concepts by treating them as whole units, whereas, the aggregated word embedding depends on the semantic similarities between individual words within the concepts. For example, we found that 'breast cancer' is closer to 'lung cancer' when aggregated word embedding is used, because the vectors are affected by occurrence of the shared word 'cancer.'

Proposed document representation and weighting scheme

In order to properly make use of the concept embedding for text clustering and visualization, in this research, a new weighting scheme is proposed ($W(C_i, d)$) to calculate the weights for the units in the vector that represents a text document. This proposed weighting scheme (shown as Eq. 4) alters the traditional tf–idf weighting scheme by considering the similarities between the concepts within the documents.

$$W(C_i, d) = \begin{cases} tf(C_i, d) \times \log \frac{|D|}{df(C_i)} + \sum_{j=1}^{M} S_{i,j} & tf(C_i, d) > 0, S_{i,j} > \theta \\ \sum_{j=1}^{N} \frac{N-(j-1)}{N} S_{i,j} & tf(C_i, d) = 0, S_{i,j} > \theta \end{cases} \quad (4)$$

$df(C_i)$: document frequency of concept C_i; $tf(C_i, d)$: frequency of concept C_i in document d; $|D|$: total number of documents in the text collection; $S_{i,j}$: the association score between concepts C_i and C_j; θ: the threshold to select the most associated concepts based on the association score; M: the number of closest concepts C_i within the same document; and N: the number of closest concepts C_i in the text collection.

The weighting scheme uses the tf–idf value to underline the occurrence of the concept in the local content. The $\sum_{j=1}^{M} S_{i,j}$ calculates the sum of association scores between the concept and the highly associated concepts that also occur within the same document. For example, if 'Essential Hypertension' and 'HTN' both occur in the document, and their association score is higher than the threshold θ, the association score is added to original tf–idf value to emphasize the co-occurrences of the concepts. By using the traditional tf–idf, if the tf value is 0, the corresponding value will be 0 for the vector representation; whereas using the proposed weighting scheme, the weight is calculated by a weighted sum of highly associated concepts that appear in the document. For example, a document does not contain the concept 'diabetes mellitus,' but contains 'diabetes.' Instead of using 0 for 'diabetes mellitus,' the similarity score between 'diabetes mellitus' and 'diabetes' is used. This proposed weighting scheme has been evaluated for text clustering and visualization in this research.

Clustering and visualization methodologies

The k means clustering algorithm (Hartigan and Wong 1979) has been successfully used in various application domains, such as text mining, computer vision, and so on (Logeswari and Premalatha 2013; Zhang et al. 2007). The aim

Table 1 Examples of top 3 most associated concepts and the association scores

Concept	Association score
Alzheimer disease	
Intact concept	
Alzheimer	0.829
Parkinson disease	0.813
Alzheimer's	0.757
Aggregated word	
Presenile alzheimer disease	0.913
Parkinson disease	0.903
Huntington disease	0.895
Multiple sclerosis	
Intact concept	
Multiple sclerosis relapsing remitting	0.661
Ms	0.633
Parkinson	0.600
Aggregated word	
Opticospinal multiple sclerosis	0.957
Progressive multiple sclerosis	0.939
Multiple sclerosis primary progressive	0.902
Cerebral amyloid angiopathy	
Intact concept	
Caa	0.601
Cerebral	0.486
Amyloid angiopathy	0.468
Aggregated word	
Senile cerebral amyloid angiopathy	0.969
Cerebral amyloid angiopathy genetic	0.965
Sporadic cerebral amyloid angiopathy	0.960
Colon cancer	
Intact concept	
Colorectal cancer	0.799
Cancer of colon	0.755
Cancer of the colon	0.725
Aggregated word	
Cancer of colon	0.980
Colon cancers	0.944
Metastatic colon cancer	0.943

of k-means algorithm is to minimize the sum of distances of each point within the cluster to the cluster center. Given $x = x_1, x_2, \ldots, x_N$ as a set of training data and $Clusters = Cluster_1, Cluster_2, \ldots, Cluster_k$ as a set of initialized k centers $(\mu_1, \mu_2, \ldots, \mu_k)$, the algorithm can be summarized as follows:

(1) Assignment of cluster centers: assign each data point x_i to the cluster $Cluster_j$ whose Euclidean distance from the cluster center is minimum of all the cluster centers.

$$Cluster_j = \{x_n : ||x_n - Cluster_j|| \leq ||x_n - Cluster_i||, \quad \forall i, i \in [1, k]\}. \tag{5}$$

(2) Update cluster centers: set the new center of each cluster to the mean of all data points belonging to that cluster.

$$\mu_i = \frac{1}{|Cluster_i|} \sum_{x_n \in Cluster_i} x_n, \quad \forall i. \tag{6}$$

(3) Repeat the previous two steps until convergence.

To visualize the cluster distributions on a two-dimensional space, the t Distributed Stochastic Neighbor Embedding (t-SNE) which was implemented by Maaten and Hinton (2008), is employed to project the high-dimensional data into two-dimensional space. The t-SNE algorithm minimizes the sum of the KL divergences of all data points in the original dimensional space and the mapping space. The cost function is given as Eq. (7).

$$\sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{7}$$

$p_{j|i}$: the conditional probability of the similarity of data point x_j to data point x_i when x_j is its neighbor in proportion to their probability density under a Gaussian centered at x_i and $q_{j|i}$: the conditional probability of the low-dimensional counterparts of x_i and x_j.

When $p_{j|i}$ and $q_{j|i}$ are equal, the value of the cost function is minimum. So, t-SNE algorithm aims to find a low-dimensional data representation that minimizes the mismatch between $p_{j|i}$ and $q_{j|i}$.

The computational cost of t-SNE is high when the original dimensionality of the data is high. To speed up the process, Principle Component Analysis (PCA) can be used to reduce the dimensionality to a lower space before t-SNE technique is applied. PCA suppresses some noise without severely distorting the distances between data points (Maaten and Hinton 2008).

Data sets

To evaluate the proposed biomedical text clustering and visualization framework, two biomedical text collections have been used. One collection is a labeled, the other is unlabeled. The details of the document collections are given as follows.

Trecgen

The Trecgen collection contains 4478 abstracts that are extracted from MEDLINE as a part of the TREC 2005 Genomics Track (Hersh et al. 2005). The abstracts are categorized into seven categories which are detailed in Table 2. After concept extraction by using MetaMap, it is discovered that about 50% of the extracted concepts of diseases contain two words, and 4.83% of the concepts have more than 5 words. Many of the concepts have low document frequency. Over 57% of the concepts have document frequency 1, and 7% of the concepts have document frequency over 10. This means very few concepts occur in more than 10 documents. And it also implies that the same diseases may be

Table 2 Summary of Trecgen

Category	No. of documents
Mad cow disease	447
Multiple sclerosis	554
Colon cancer	567
Alzheimer's disease	1201
Cerebral amyloid angiopathy	482
Parkinson's disease	769
Breast cancer	458
Total	4478

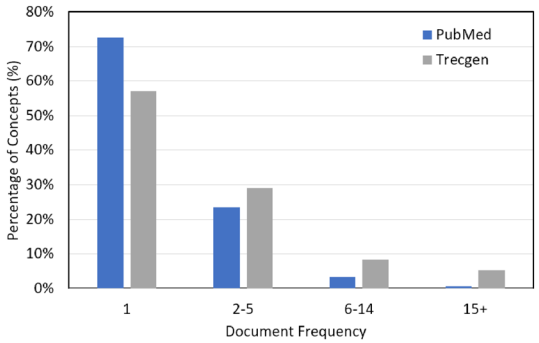

Fig. 4 Document frequency of the concepts

represented differently in different documents. Figure 4 shows the document frequency of the extracted concepts in this text collection.

PubMed OA

The PubMed Central-Open Access (PubMed OA) data set has been widely used in many research projects for biomedical text clustering and classification (Zhang et al. 2007; Zhu et al. 2009). Different from Trecgen, PubMed OA is an unlabeled (un-categorized) collection which contains over 1 million articles from the database—PubMed Central. In this research, 600 abstracts were randomly selected from journals whose names begin with letter 'A' or 'B.' The number of selected articles from each journal is shown in Table 3.

After concept extraction by using MetaMap, it is discovered that around 20% of the concepts are single word concepts, 50% of them contain two words, and around 9% of them contain four or more words. Figure 4 shows that many of the concepts have low document frequency. Over 72% of the concepts have document frequency 1, and less than 5% of the concepts have document frequency over 6.

Table 3 Summary of PubMed OA

Name of journal	No. of documents
Augmentative and Alternative Communication	2
American Journal of Physiology Endocrinology and Metabolism	11
Biological Trace Element Research	31
Ancient Science of Life	3
Allergy and Asthma Proceedings	28
Brain and Language	1
Bone Marrow Research	1
BoneKEy Reports	4
Annals of Rehabilitation Medicine	323
Aphasiology	3
Anesthesia, Essays and Researches	135
Bioinformatics and Biology Insights	45
American Journal of Hypertension	13
Total	600

Clustering evaluation metrics

To identify the optimum number of clusters, the clustering results are evaluated against clustering evaluation metrics. Typically, there are two types of evaluation metrics: internal evaluation and external evaluation. The internal evaluation is to formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity. Davies–Bouldin index (Davies and Bouldin 1979) is one of the internal evaluation metrics. The external evaluation metrics are based on the interest of an application, such as categorization. F-measure is one of the external evaluation metrics. The Davies–Bouldin index and F-measure are briefly described as follows:

- Davies–Bouldin index: The calculation of the Davies–Bouldin index is shown in Eq. (8), where $\sigma_{Cluster_i}$ is the standard deviation of the distance of samples in a cluster $Cluster_i$ to the respective cluster centroid; $d(Cluster_i, Cluster_j)$ is the Euclidean distance between centroids of cluster $Cluster_i$ and $Cluster_j$; and $|Clusters|$ is the total number of clusters. The more distinct the clusters are from each other, the lower the DB index value is

$$DBindex = \frac{1}{|Clusters|} \sum_{i=1}^{|Clusters|} \max_{j,j\neq i} \frac{\sigma_{Cluster_i} + \sigma_{Cluster_j}}{d(Cluster_i, Cluster_j)}. \tag{8}$$

- F-measure: If each cluster is labeled by the category of the majority of the data points within the cluster, the F-measure is computed as Eq. (9), in which tp, fn, and fp stand for true positive, false negative, and false positive.

$$F\text{-}measure = \frac{tp}{2 \times tp + fn + fp}. \tag{9}$$

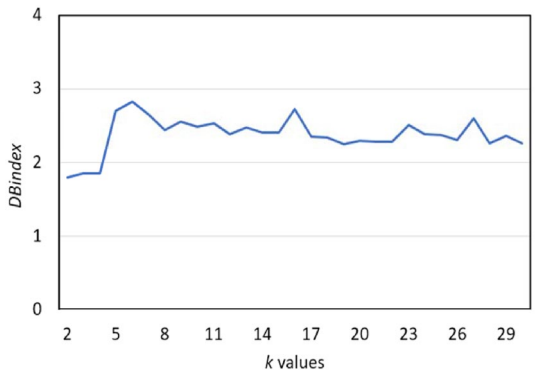

Fig. 5 Trecgen—*DBindex*—baseline TF–IDF

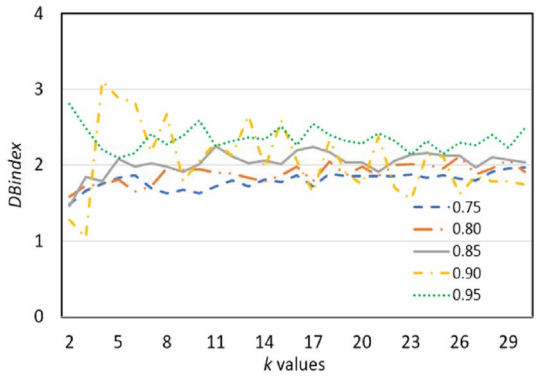

Fig. 6 Trecgen—*DBindex*—aggregated word embedding

Clustering performance analysis and visualization

The threshold θ in Eq. (4) determines the selected associated concepts to calculate the weights for document representations. In this research, we fully evaluated the θ from 0.75 to 0.95 for both aggregated word embedding and intact concept embedding. Different k values (from 2 to 30) for the k means algorithm have also been explored for both data sets. The visualizations are performed based on selected k values and θ values. The detailed clustering performance and visualization on each data set are provided in the following sub-sections.

Trecgen

Figures 5, 6, and 7 show the returned values of the *DBindex* for baseline tf–idf, proposed weighting scheme based on aggregated word embedding and intact concept embedding for different k values. The returned *DBindex* values show that the proposed weighting scheme based on aggregated word embedding works better than the other two.

Figures 8, 9, and 10 show the returned values of the F-measures for different k values of the three weighting schemes. Since the F-measure evaluates the clustering result

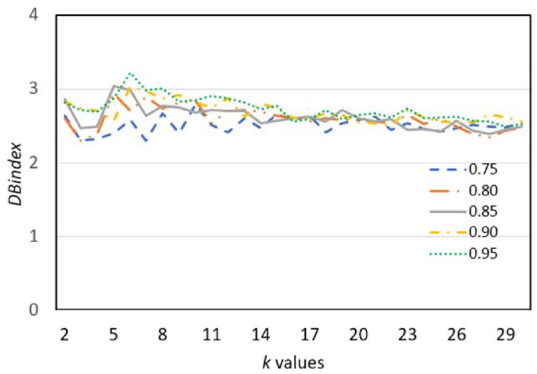

Fig. 7 Trecgen—*DBindex*—intact concept embedding

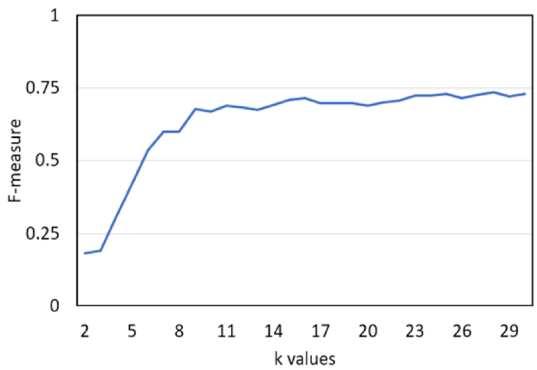

Fig. 8 Trecgen—*F-measure*—baseline TF–IDF

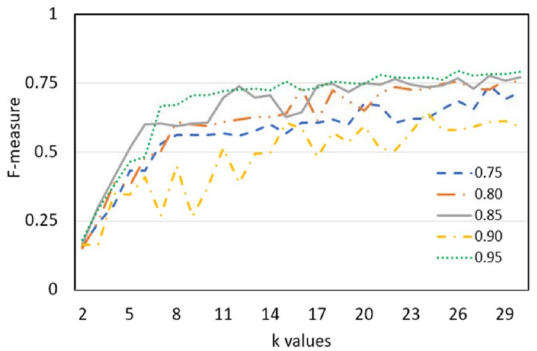

Fig. 9 Trecgen—*F-measure*—aggregated word embedding

according to the true categories of the documents, the higher the F-measure is, the better clustering performance. Different from the returned *DBindex* values, the returned F-measures show that the proposed weighting scheme based on intact concept embedding performs better than the other two. The F-measures of all three show that the values start to have minor changes when k is larger than 7 which is the number of categories of this data collection.

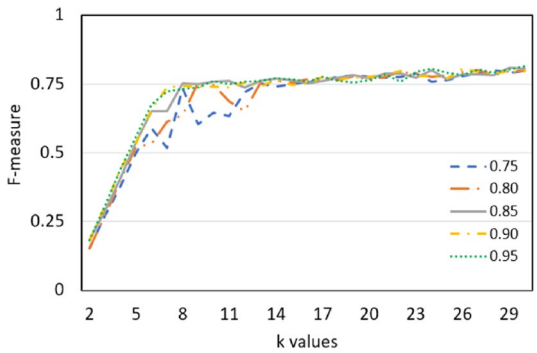

Fig. 10 Trecgen—*F-measure*—intact concept embedding

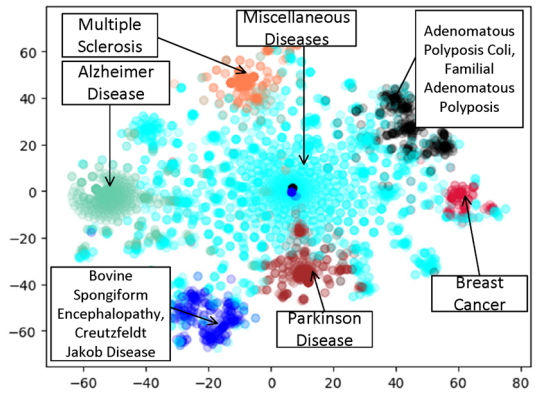

Fig. 11 Trecgen—baseline TF–IDF—$k = 7$

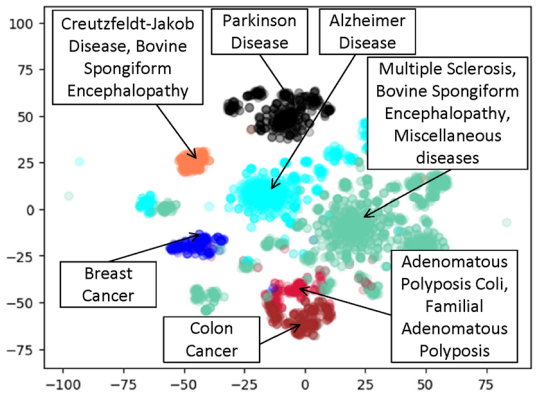

Fig. 12 Trecgen—aggregated word embedding—$k = 7, \theta = 0.85$

To visualize the distribution of clusters, PCA is used to reduce the original dimension to 900, then *t*-SNE algorithm is applied to visualize the clusters on the two-dimensional space. The clusters are labeled by the concept(s) that have top document frequency within the clusters. Figures 11, 12, and 13 show the visualization of the three weighting schemes when $k = 7$. The baseline tf–idf shows 7 clusters that match six categories

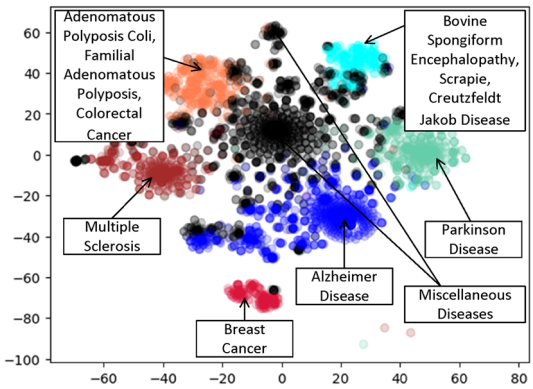

Fig. 13 Trecgen—intact concept embedding—$k = 7, \theta = 0.90$

in the original document collection. The cluster 'Miscellaneous disease' contains many different diseases and has overlapping with all other clusters. The category 'Cerebral Amyloid Angiopathy' is not notable in the visualization. After investigation, it is found that most of the documents in the 'Cerebral Amyloid Angiopathy' category are in either 'Alzheimer disease' or 'Miscellaneous disease' cluster. Figure 12 shows 7 clusters using the proposed weighting scheme based on the aggregated word embedding when θ is set to 0.85. Comparing to the visualization of baseline tf–idf, it is apparent that the clusters based on the aggregated word embedding are well separated with fewer overlaps. Documents in category 'Colon Cancer' are included in two clusters that are next to each other. The documents within category 'Cerebral Amyloid Angiopathy' are distributed across the clusters of 'Alzheimer disease' and 'Parkinson disease.' It is also noticed that the cluster on the right (light green) mixed 'Multiple Sclerosis' and 'Bovine Spongiform Encephalopathy' which is a kind of mad cow disease. This explains that the F-measure based on the aggregated word embedding is low then k is 7. Figure 13 shows 7 clusters using the proposed weighting scheme based on the intact concept embedding when θ is set to 0.90. The separation of the clusters is not as obvious as that based on the aggregated word embedding. This explains that the *DBindex* is higher than that based on the aggregated word.

PubMed OA

Since PubMed OA is a collection without defined categories for the documents, only internal evaluation *DBindex* is used to evaluate the clustering performance. Figures 14, 15, and 16 show the *DBindex* values for baseline tf–idf, proposed weighting scheme based on aggregated word embedding, and intact concept embedding for different k values. The *DBindex* values show that the proposed weighting schemes based on aggregated word embedding and intact concept embedding work much better than the baseline tf–idf. The *DBindex* values based on the aggregated word embedding are slightly lower than those based on the intact concept embedding. Although the *DBindex* decreases along with the increasing of the k, it does not decrease significantly decrease when k is larger than 14.

To visualize the distribution of clusters, PCA is used to reduce the original dimension to 300, then t-SNE algorithm is applied to visualize the clusters on the two-dimensional

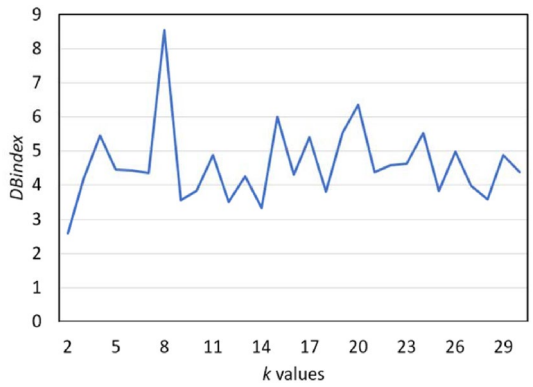

Fig. 14 PubMed OA—*DBindex*—baseline TF–IDF

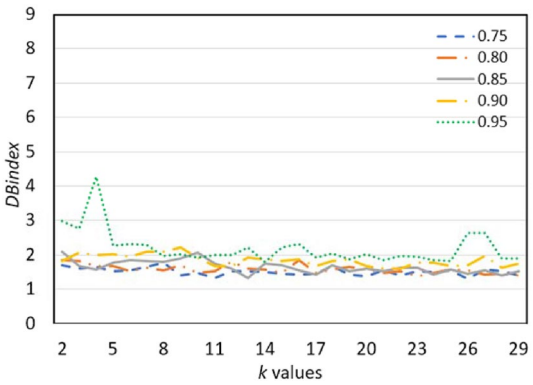

Fig. 15 PubMed OA—*DBindex*—aggregated word embedding

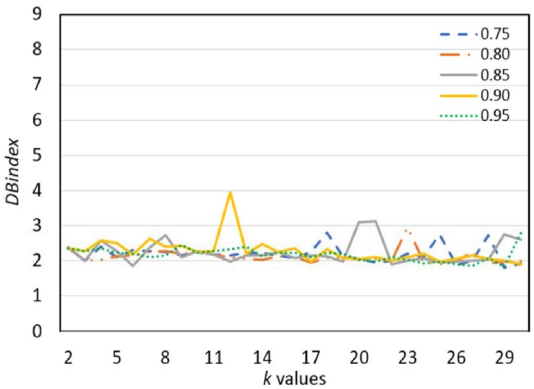

Fig. 16 PubMed OA—*DBindex*—intact concept embedding

space. The clustering visualization of $k = 14$ based on the baseline tf–idf is given in Fig. 17. Apparently, the baseline tf–idf does not work well on generating document clusters. The largest cluster (black), and some small clusters (light blue, blue, red, and brown) all contain documents of many diseases that include 'stroke,' different types of cancers, and other cardiac and neurological diseases. Hence, they are all labeled as

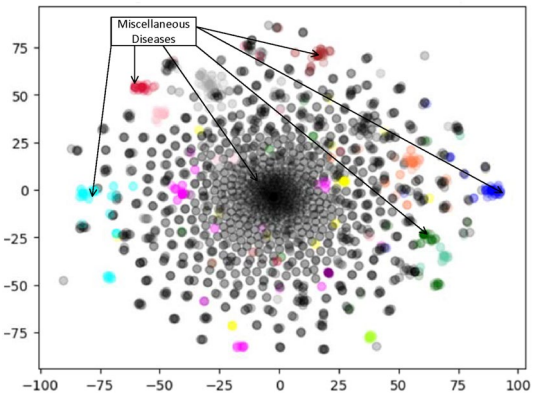

Fig. 17 PubMed OA—baseline TF–IDF—$k = 14$

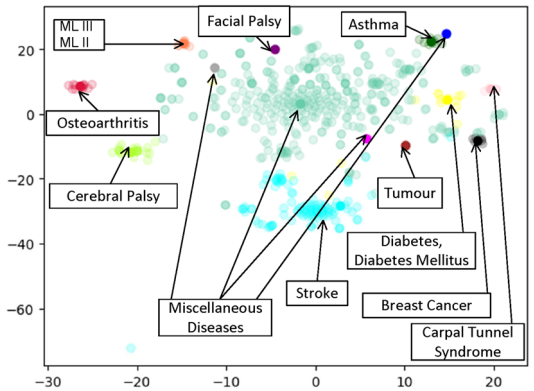

Fig. 18 PubMed OA—aggregated word embedding—$k = 14, \theta = 0.85$

'Miscellaneous Diseases.' There is no cluster that has high document frequency of a specific disease concept.

Figures 18 and 19 show the clustering visualization of selected θ based on the aggregated word embedding and intact concept embedding when *DBindex* are low and $k = 14$. Comparing to the baseline tf–idf, the clusters shown in these figures are well formed and separated from each other. It is found that the densities of some clusters given in Fig. 18 is slightly higher than those in Fig. 19. This explains that when $k = 14$, the *DBindex* value based on the aggregated word embedding is lower than that based on the intact concept embedding. We then label the clusters by the concept(s) that have top frequent document frequency. It is found that the intact concept embedding can identify more clusters of specific types of diseases, such as 'Obesity,' 'Chronic Low Back Pain,' and 'Hypertension.' We also notice that some clusters contain some concepts that are highly related. For example, the 'Breast Cancer' cluster contains many instances of 'Lymphoedema.' The 'Hypertension' cluster contains documents describing hypertension, hypoglycemia, and heart blockage. By analyzing the clusters based on the aggregated word embedding, we found that some cluster, such as 'Tumor,' contains documents that are under a category. We hypothesize that the reason is that many of the concepts have

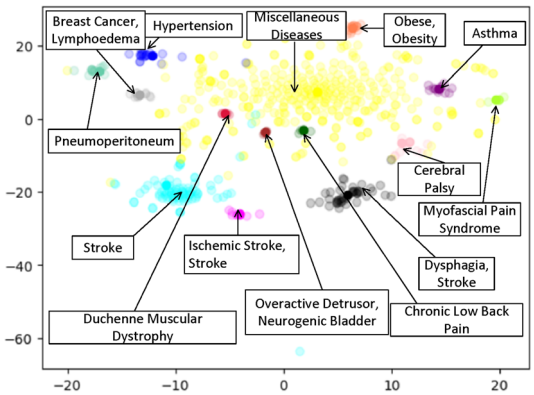

Fig. 19 PubMed OA—intact concept embedding—$k = 14, \theta = 0.85$

some common word(s). Actually, it is found that 'brain tumor' and 'tumor in breast' are all in cluster 'Tumor.'

Discussion

The clustering evaluation and visualization show that both aggregated word embedding and intact concept embedding-based weighting schemes perform better than baseline tf–idf weighting scheme. The concept embedding captures the associations between the concepts, and by making use of the concept associations, documents can be well represented for the task of biomedical document clustering. The returned *DBindex* values based on aggregated word embedding approach are slightly better than those based on the intact concept embedding approach. However, since *DBindex* evaluates the clustering performance based on high intra-cluster similarity and low inter-cluster similarity without considering the content within the clusters, low *DBindex* value might not reflect high content similarity within the clusters; whereas the external evaluation F-measure considers the content wise intra-cluster similarity. This explains that the returned F-measures on the Trecgen show that the intact concept embedding works better than the other two. Based on the visualizations, the intact concept embedding can better identify more clusters of specific diseases; whereas, the aggregated word embedding can identify the clusters of higher level categories of diseases, such as cancer or tumor. So, it is worth to investigate the integration of these two concept embedding for clustering and visualization in the future.

Conclusion and future work

In this paper, a framework for biomedical document clustering and visualization based on concept embedding of diseases is proposed and evaluated. The concept embedding is learned through neural networks. The first concept embedding is based on aggregating word embedding, and the second is intact concept embedding. Both concept embedding generation approaches can capture the association of the concepts based on the content of the training text collection. A new weighting scheme is proposed to use the concept embedding and compared against baseline tf–idf for text clustering and visualization.

The results demonstrate that the proposed weighting scheme using the concept embedding achieve better performance than the baseline tf–idf.

Potential future work includes extending this framework to biomedical document clustering by including concept embedding of other types of medical concepts, such as symptoms, treatments, and so on. On the other hand, other clustering methods and hierarchical clustering architecture can be explored for the clustering and the visualization of larger text collection.

Authors' contributions
The authors discussed the problem and the solutions were proposed all together. All authors participated in drafting and revising the manuscript. Both authors read and approved the final manuscript.

Author details
[1] Department of Computer Information Technology, IUPUI, Indianapolis, USA. [2] Department of Electrical and Computer Engineering, IUPUI, Indianapolis, USA.

Acknowledgements
This project is supported by IUPUI Enhanced Mentoring Program with Opportunities for Ways to Excel in Research (EMPOWER).

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

Consent for publication
All authors consent to publish this work.

Ethics approval and consent to participate
Not applicable.

Funding
There was no external funding for this study. Internal funding support was written in acknowledgement.

Publisher's Note

Received: 18 August 2018 Accepted: 20 October 2018
Published online: 01 November 2018

References

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Patt Anal Mach Intell 2:224–227
Fact Sheet—UMLS Metathesaurus. https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
Globerson A, Chechik G, Pereira F, Tishby N (2007) Euclidean embedding of co-occurrence data. J Mach Learn Res 8:2265–2295
Gorg C, Tipney H, Verspoor K, Baumgartner WK, Cohen KB, Stasko J, Hunter LE (2010) Visualization and language processing for supporting analysis across the biomedical literature. In: International conference on knowledge-based and intelligent information and engineering systems proceedings, pp 420–429
Gu J, Feng W, Zeng J, Mamitsuka H, Zhu S (2013) Efficient semisupervised medline document clustering with mesh-semantic and global-content constraints. IEEE Trans Cybern 43(4):1265–1276
Hartigan JA, Wong MA (1979) Algorithm as 136: a k-means clustering algorithm. J Royal Stat Soc. 28(1):100–108
Hersh, W, Cohen, A, Yang, J, Bhupatiraju RT, Roberts P, Hearst M (2005) Trec 2005 genomics track overview. NIST Special Publication 500-266: The Fourteenth Text REtrieval conference proceedings
Kim HK, Cho S (2017) Bag-of-concepts : comprehending document representation through clustering words in distributed representation. Neurocomputing
Levy O, Goldberg Y (2014) Neural word embedding as implicit matrix factorization. In: Advances in neural information processing systems, pp 2177–2185
Logeswari S, Premalatha K (2013) Biomedical document clustering using ontology based concept weight. In: International conference on computer communication and informatics proceedings, pp 1–4, https://doi.org/10.1109/ICCCI.2013.6466273
Lvd Maaten, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res. 9:2579–2605
MEDLINE/PubMed Resource Guide. https://www.nlm.nih.gov/bsd/pmresources.html
MetaMap—a tool for recognizing UMLS concepts in text. https://metamap.nlm.nih.gov/
Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: International conference on neural information processing systems, pp 3111–3119

Moen S, Ananiadou TSS (2013) Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th international symposium on languages in biology and medicine, Tokyo, Japan, pp 39–43
Pearson K (2008) Liii on lines and planes of closest fit to systems of points in space. London Edinburgh Dublin Philos Magaz J Sci 2(11):559–572. https://doi.org/10.1080/14786440109462720
PubMed Open Access Subset. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/
Shah S, Luo X (2018) Comparison of deep learning based concept representations for biomedical document clustering. In: IEEE EMBS international conference on biomedical & health informatics (BHI), pp 349–352. IEEE, New York
Tulkens S, Daelemans W (2016) Using distributed representations to disambiguate biomedical and clinical concepts. arXiv preprint arXiv:1608.05605
Van Rijsbergen C (1979) Information retrieval. dept. of computer science, university of glasgow https://citeseer.ist.psu.edu/vanrijsbergen79information.html
Yoo I, Hu X, Song I-Y (2006) A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. In: First international workshop on text mining in bioinformatics proceedings, pp 84–89
Zhang X, Jing L, Hu X, Ng M, Zhou X (2007) A comparative study of ontology based term similarity measure on pubmed document clustering. In: International conference on database systems for advanced applications proceedings, pp 115–126
Zhu Y, Yan E, Wang F (2017) Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. BMC Med Inform Decis Making 17:95–103
Zhu S, Zeng J, Mamitsuka H (2009) Enhancing medline document clustering by incorporating mesh semantic similarity. Bioinformatics 25(15):1944–1951