Comparison of Deep Learning based Concept Representations for Biomedical Document Clustering

Setu Shah[1] and Xiao Luo[1]

Abstract— **In this research, document representations based on distributed representations of the concepts along with new weighting schemes for the documents are explored. The baseline weighting scheme is the traditional Term Frequency-Inverse Document Frequency (TF-IDF) of the concepts, whereas, the other two newly proposed ones consider both local content using the TF-IDF and associations between concepts. The distributed representations of the concepts are measured using a deep learning algorithm. The evaluation of the proposed document representations is based on the k-means clustering results. The results show that document representation based on TF-IDF in combination with the term based distributed representations for concepts outperforms the other two based on the returned evaluation metrics - F1-measure (80.21%) and Purity (77.1%).**

I. Introduction

Active research in the medical and biomedical domain has generated pervasive documents and articles. There is a continuous need for development of techniques to discover, search, access and share knowledge from these documents and articles. Biomedical document clustering is different from the general text document clustering task; one medical concept of disease might be represented in different forms, and some medical concepts of diseases might be highly correlated. For example, 'Type 2 Diabetes' is the same concept of disease as 'Diabetes Mellitus Type 2'. 'Hypertension' might co-occur often with 'Stroke'. In order to capture the semantic similarities between words or phrases, previous research on document representation reforming [1] [2] often use existing ontology such as MeSH or WordNet to identify the semantic relationships. However, ontology doesn't reflect the co-occurrences of medical concepts. Moreover, generating and updating the ontology requires substantial human resources and time.

In this research, document representations based on distributed representations of the concepts and proposed weighting scheme are explored. Specifically, concepts of diseases are extracted from the biomedical document corpus using UMLS MetaMap to construct the vectors of documents. Then, two deep learning based distributed representations for the extracted concepts of diseases are investigated. One is built upon single word distributed representations, the other is built upon distributed representations for terms which include one or more than one word. The k-means clustering algorithm is employed for the stage of clustering. In this research, a series of k values have been explored to fully evaluate the results. The clustering results have been evaluated through evaluation metrics - Purity and F1-measure. The overall results demonstrate that the document representation based on TF-IDF in combination with the term based distributed representations outperforms the traditional TF-IDF and TF-IDF in combination with the word based distributed representations.

The rest of the paper is organized as follows. In section II, related work is described. Section III provides the details of the concept extraction and deep learning based distributed representation. Section IV describes the calculation of the similarity between the concepts. Document representation and weighting scheme are presented in Section V. Experimental data set, clustering evaluation, and discussion is given in section VI. Section VII concludes this research and discusses potential future work.

II. Related Work

A lot of research has been done in biomedical document clustering in the past decades. Zhang et al. [2] reviewed three different ontology based term similarity measurements: path-based, information content-based and feature-based, and then proposed their own similarity measurement and term reweighting scheme. The k-means algorithm is used for document clustering. Based on the results comparison, some of them are slightly worse than the single word based weighting scheme which gained 75% F1-measure. Logeswari et al. [1] proposed a concept weighting scheme based on the MeSH ontology and tri-gram extraction to extract concepts from the text corpus. The semantic relationship between tri-grams are weighted through a heuristic weight assignment of four predefined semantic relationships. The k-means clustering algorithm results show that concept representation was better than word representation. The best returned purity evaluation was below 60%. Gu et al. [3] proposed a concept similarity measurement by using a linear combination of multiple similarity measurements based on the MeSH ontology and the local content which includes TF-IDF weighting and coefficient calculation between related document sets. A semi-supervised clustering algorithm was employed at the stage of document clustering. Drakopoulos et al. [4] compared three different document representations for biomedical document clustering. They found with the increase in the size of the document set, the performance decreased. The performance of using whole document set and tensor based document representation gained 56.35% on the F1-measure.

To the best of authors' knowledge, this research is the first to compare deep learning based concept representations for

[1] Setu Shah and Xiao Luo are with Purdue School of Engineering Technology, IUPUI, 799 W. Michigan Street, Indianapolis, IN 46202, USA setshah@iupui.edu, luo25@iupui.edu

biomedical document clustering based upon traditional TF-IDF and its combination with concepts' associations based on their distributed representations.

III. Concepts Extraction and Distributed Representation

In this work, clustering is based on the concepts of diseases that are mentioned in the documents. To extract the concepts of diseases from the documents, UMLS MetaMap [5] is used. UMLS MetaMap is a natural language processing tool to map the phrases or terms in the text document to different semantic types. In this research, if a term or phrase is mapped to semantic types 'Disease or Syndrome' or 'Neoplastic Process', the corresponding lexical phrase identified by MetaMap is extracted as a concept of disease and used for document representation construction.

The distributed representation of words [6] based on the deep learning algorithm (Recurrent Neural Network) has drawn attention in the areas of natural language processing and machine learning [7] [8] [9]. The learned distributed representation of words preserve the distances between words, so that the words that have semantic and syntactic associations in the raw text corpus are located in close proximity to one another. The dimension of the vector created depends on the number of neurons in the hidden layer of the neural network.

In this research, two Recurrent Neural Networks (RNNs) based on the skip-gram model are trained to create the distributed representation (DisV) for the concepts of the diseases. One is trained by inputting single words, the other one is trained based on the terms presenting the concepts of diseases and non-disease related words in the raw document set.

- Word based concept representation
  The input to the RNN are single words of the training set which includes PubMed Central's Non-Commercial Open Access database [10] and a subset of documents from MEDLINE [11]. The output are vectors representing words - $(wv_1, wv_2, \ldots, wv_m)$. If a concept of disease includes multiple words, the concept vector is generated by aggregating the vectors representing words within the concept, as shown in Equation 1. For example, for the disease 'diabetes mellitus', the vectors representing 'diabetes' and 'mellitus' are aggregated by adding them together.

  $$DisV = \sum_{i=1}^{M}(wv_{i1}, wv_{i2}, \ldots, wv_{im}) \qquad (1)$$

  M: the total number of words in a concept $DisV$.
- Term based concept representation
  The input data set is the same as word based concept representation. Instead of creating vectors for single words and aggregating them for concept representation, each extracted concept is treated as a word. So, the input to the RNN model includes terms presenting the concepts of disease and non-disease words within the training set. At the end of the training process, each concept of disease is represented as a vector directly, as shown in Equation 2.

$$DisV = (disv_1, disv_2, \ldots, disv_m) \qquad (2)$$

In this research, the dimension of the vectors (m) which is the number of neurons in the hidden layer of the RNNs are all set to 300.

IV. Disease Association Matrix Construction

Since each concept of disease is represented by a vector (Equation 2), the similarity or the association between the concepts can be measured through a distance calculation between the vectors. All the association scores between any two diseases can be stored in a matrix.

Given a total of L concepts of diseases extracted from the raw biomedical document corpus, the association scores are stored in the matrix S as presented in Equation 3. Each entry $s_{i,j}$ in the matrix S represents the association score between concept $DisV_i$ and $DisV_j$. Cosine distance (Equation4) is used to calculate the association scores.

$$S_{L,L} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,L} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ s_{L,1} & s_{L,2} & \cdots & s_{L,L} \end{pmatrix} \qquad (3)$$

$$S_{i,j} = \frac{DisV_i\dot{}DisV_j}{||DisV_i||_2||DisV_j||_2} \qquad (4)$$

Table I provides some examples of concepts of diseases and the their top 3 closest concepts based on the association scores computed from two different learning models for two concept representations: word based and term based. Both concept representations capture the associations between two concepts of diseases. The association scores of the term based representation are lower than the word based representation in general. The reason is that the term based representation measures the similarity of the semantic meanings of the concepts more closely, whereas, the word-based representation is heavily dependent on the similarity of individual words within the concepts. Thresholds are applied to identify the most similar concepts. For the word based representation, if the association score is above 0.9, it implies that two concepts are highly associated or similar. Whereas, for the term based representation, 0.75 is used as a threshold to identify the most similar concepts. In section V, these two thresholds are used to select the most similar concepts for document representation.

V. Document Representation and Weighting Scheme

In this research, a new weighting scheme is proposed ($W_{Dis_{i,d}}$) to calculate the weight for each concept ($Dis_{i,d}$) in the vector representation. It alters the traditional TF-IDF weighting scheme by considering the similarities between the concepts and co-occurrences of the concepts within the

TABLE I

EXAMPLES OF TOP 3 CLOSEST CONCEPTS AND THE ASSOCIATION SCORES

Concept	**Term-based Representation**	**Term-based Association Score**	**Word-based Representation**	**Word-based Association Score**
alzheimer disease	alzheimer	0.829	presenile alzheimer disease	0.913
	parkinson disease	0.813	parkinson disease	0.903
	alzheimer's	0.757	huntington disease	0.895
multiple sclerosis	multiple sclerosis relapsing remitting	0.661	opticospinal multiple sclerosis	0.957
	ms	0.633	progressive multiple sclerosis	0.939
	parkinson	0.600	multiple sclerosis primary progressive	0.902
cerebral amyloid angiopathy	caa	0.601	senile cerebral amyloid angiopathy	0.969
	cerebral	0.486	cerebral amyloid angiopathy genetic	0.965
	amyloid angiopathy	0.468	sporadic cerebral amyloid angiopathy	0.960
colon cancer	colorectal cancer	0.799	cancer of colon	0.980
	cancer of colon	0.755	colon cancers	0.944
	cancer of the colon	0.725	metastatic colon cancer	0.943

documents. The weighting scheme is calculated as equation 5:

$$W_{Dis_{i,d}} = \begin{cases} tf_{Dis_{i,d}} \times \log \frac{|D|}{df_{Dis_i}} + \sum_{j=1}^{M} S_{i,j} & tf_{Dis_{i,d}} > 0 \\ \sum_{j=1}^{N} \frac{N-(j-1)}{N} S_{i,j} & tf_{Dis_{i,d}} = 0 \end{cases} \tag{5}$$

df_{Dis_i}: document frequency of concept Dis_i

$tf_{Dis_{i,d}}$: frequency of concept Dis_i in document d

$|D|$: total number of documents in the corpus

$S_{i,j}$: the association between Dis_i and concept Dis_j; $S_{i,j} >= 0.75$ for term based representation; $S_{i,j} >= 0.9$ for word based representation

M: the number of closest concepts within the same document

N: the number of closest concepts of Dis_i in the corpus

If a concept occurs in a document, the weighting scheme uses the traditional 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 occurred concept $Dis_{i,d}$ and the highly associated concepts that also occur within the document. For example, if 'Essential Hypertension' and 'HTN' both occur in the document, and their association score is more than the thresholds, their association scores will be added to original TF-IDF value to emphasize the occurrences of the concepts. If a concept does not occur in the document, the weight is calculated by a weighted sum of the highly similar concepts that appear in the document. For example, 'diabetes' occurs in one document, but 'diabetes mellitus' occurs in another document. By using the traditional TF-IDF weighting scheme, their values would be 0 for these documents. However, by using the proposed weighting scheme, for the document that does not contain the concept 'diabetes mellitus', instead of using 0, the similarity score between 'diabetes mellitus' and other concepts, e.g. 'diabetes', that appear in the document is used. This weighting scheme has been explored based on the association scores of two kinds of concepts representations: word based and term based. We refer them as 'TF-IDF + Word-based' and 'TF-IDF + Term-based' in the experimental section below.

TABLE II

OVERVIEW OF THE DATA SET

Disease	**# of documents**
Multiple Sclerosis	554
Mad Cow Disease	447
Alzheimer's Disease	1201
Colon Cancer	567
Parkinson's Disease	769
Cerebral Amyloid Angiopathy	482
Breast Cancer	458
Total	4478

VI. EXPERIMENTAL RESULTS AND DISCUSSION

A. Data Set

To evaluate the proposed biomedical document clustering framework, a subset of large biomedical document collections from MEDLINE has been used [12]. The categories and the corresponding number of documents in this data set are detailed in Table II. In this research, only content in the 'Title' and 'Abstract' sections of these documents are used for document clustering. After extracting the concepts of diseases from the document, the analysis of the document frequencies of the concepts shows that over 57% of the concepts have document frequency 1, and only 7% of the concepts have document frequency over 10. That means not many concepts of diseases occur in more than 3 documents. If traditional TF-IDF weighting scheme is used, majority of the entries in the vectors are 0. Figure 1 shows the distribution of the concepts based on the number of words in each concept. Majority of the concepts of diseases contain more than one word.

B. Evaluation Metrics and Results

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. The external evaluation is based on the interest of an application, such as categorization. Since, the documents within the experimental data set are labelled with categories, external evaluations: F1-measure [14] and Purity [15] are used to evaluate the results.

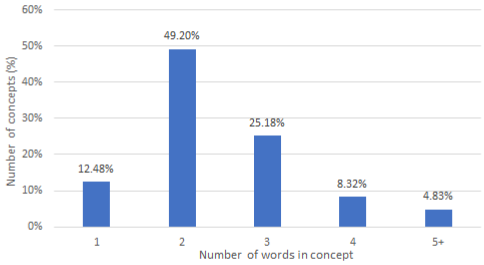

Fig. 1. Distribution of the concepts based on the number of words.

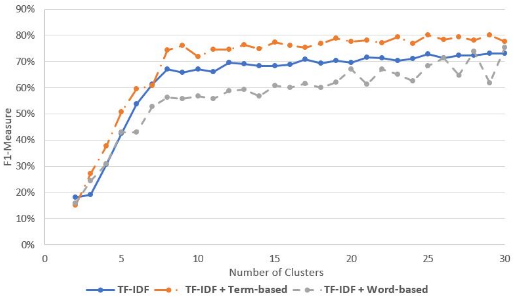

Fig. 2. F1-measures over the number of clusters (k value)

If each cluster in the set of clusters $C_1, \ldots, C_j$ is labeled by the category of the majority of the data points within the cluster, the F1-measure computed as Equation 6.

$$F1 = \frac{true\ positive}{2 \times true\ positive + false\ negative + false\ positive} \tag{6}$$

Purity is computed as Equation 7. N is the total number of data points. The set of clusters $C_1, \ldots, C_j$ that are generated by the clustering algorithm, and L the set of true categories $L_1, \ldots, L_k$.

$$Purity = \frac{1}{|N|} \sum_k \max_j |L_k \cap \ C_j)| \tag{7}$$

Figure 2 and Figure 3 show the F1-measures and Purity of the clustering results of k values from 2 to 30. The F1-measure and Purity values start to stable without significant increasing when k is larger than 8. Both evaluations show that for most of the k values, TF-IDF combined with term based concept representation performs better than the other two. It achieves the the best F1-measure - 80.2% and purity value - 77.1% when k is 29. The results on each individual category show that term based concept representation performs better on category ‘Mad Cow Disease’, ‘Cerebral Amyloid Angiopathy’, ‘Breast Cancer’ and ‘Alzheimer’s Disease’. Whereas, the word based concept representation performs better on the category ‘Multiple Sclerosis’. The reason could be that each word of the disease ‘Multiple’ and ‘Sclerosis’ is unique compared to other diseases.

VII. Conclusion and Future Work

In this paper, through the document clustering, we show that document representation based on the weighting scheme of TF-IDF combined with term based distributed concept representation performs better than the other two based on returned Purity and F1-measure values. Potential future work includes evaluating clustering based on combinations of different concepts that include concepts of symptoms and treatments and so on.

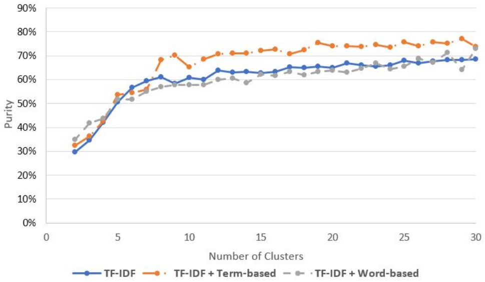

Fig. 3. Purity over the number of clusters (k value)

References

[1] S. Logeswari and K. Premalatha, “Biomedical document clustering using ontology based concept weight,” in *International Conference on Computer Communication and Informatics Proceedings*, Jan. 2013, pp. 1–4.

[2] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou, “A comparative study of ontology based term similarity measure on pubmed document clustering,” in *International Conference on Database Systems for Advanced Applications Proceedings*, 2007, pp. 115–126.

[3] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu, “Efficient semisupervised medline document clustering with mesh-semantic and global-content constraints,” *IEEE Transactions on Cybernetics*, vol. 43, no. 4, pp. 1265–1276, August 2013.

[4] G. Drakopoulos, A. Kanavos, S. S. I. Karydis, and A. G. Vrahatis, “Tensor-based semantically-aware topic clustering of biomedical documents,” *Computation*, vol. 5, pp. 34–50, 2017.

[5] *MetaMap - A Tool For Recognizing UMLS Concepts in Text*, https://metamap.nlm.nih.gov/.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *International Conference on Neural Information Processing Systems*, 2013, pp. 3111–3119.

[7] S. Š. Stéphan Tulkens and W. Daelemans, “Using distributed representations to disambiguate biomedical and clinical concepts,” *arXiv preprint arXiv:1608.05605*, 2016.

[8] H. K. Han Kyul Kim and S. Cho, “Bag-of-concepts Comprehending document representation through clustering words in distributed representation,” *Neurocomputing*, 2017.

[9] Y. Zhu, E. Yan, and F. Wang, “Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec,” *BMC Medical Informatics and Decision Making*, vol. 17, pp. 95–103, 2017.

[10] *Open Access Subset*, https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

[11] *MEDLINE/PubMed Resource Guide*, https://www.nlm.nih.gov/bsd/pmresources.html.

[12] W. Hersh, A. Cohen, J. Yang, R. T. Bhupatiraju, P. Roberts, and M. Hearst, “Trec 2005 genomics track overview,” 2005.

[13] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[14] Y. Zhao and G. Karypis, “Criterion functions for document clustering: Experiments and analysis,” *Technical Report TR 0140, Department of Computer Science, University of Minnesota*, 2001.

[15] E. Amigo, J. Gonzalo, J. Artiles, and V. Felisa, “A comparison of extrinsic clustering evaluation metrics based on formal constraints,” *Information retrieval*, vol. 12, no. 4, pp. 461–486, 2009.