Interpretable Word Embeddings For Medical Domain

Kishlay Jha*, Yaqing Wang*, Guangxu Xun, Aidong Zhang
Department of Computer Science and Engineering
State University of New York at Buffalo, NY, USA
Email: {kishlayj, yaqingwa, guangxux, azhang}@buffalo.edu

***Abstract*—Word embeddings are finding their increasing application in a variety of biomedical Natural Language Processing (bioNLP) tasks, ranging from drug discovery to automated disease diagnosis. While these word embeddings in their entirety have shown meaningful syntactic and semantic regularities, however, the meaning of individual dimensions remains elusive. This becomes problematic both in general and particularly in sensitive domains such as bio-medicine, wherein, the interpretability of results is crucial to its widespread adoption. To address this issue, in this study, we aim to improve the interpretability of pre-trained word embeddings generated from a text corpora, and in doing so provide a systematic approach to formalize the problem. More specifically, we exploit the rich categorical knowledge present in the biomedical domain, and propose to learn a transformation matrix that transforms the input embeddings to a new space where they are both interpretable and retain their original expressive features. Experiments conducted on the largest available biomedical corpus suggests that the model is capable of performing interpretability that resembles closely to the human-level intuition.**

***Index Terms*—word embeddings, interpretability, biomedicine**

I. Introduction

Modelling the lexical semantics behind a word has acquired significant interest in the recent years [1]–[3]. As a consequence of advances made in the research area of deep learning, more recently, practitioners in the community have applied neural network inspired language models (commonly known as word embedding models [2]) to model the latent structure present in the text, and produced more nuanced form of word representations. Simply put, these word embeddings models learn to generate dense, continuous, low-dimensional vectors representation of words from raw, unannotated corpora in a completely unsupervised manner. Such succinct form of representation thesedays have become the "de-facto" word representation for a multitude of downstream bioNLP tasks such as disease diagnosis [4], drug re-purposing [5] and hypotheses generation [6], [7].

Despite their considerable success and widespread adoption, a drawback of these word embedding models lie in their inability to provide meaningful interpretation of the individual embedding dimensions. This is problematic because even though we can comprehend the underlying mathematical principles of such models, it is still important to understand what exactly do these dimensions signify? What kind of properties are being (and *not* being) captured by these dimensions? As a simple illustration, consider the example of medical concepts "Insulin" and "Diabetes mellitus" shown in Figure 1. As it can be observed, the current word embedding models can capture the semantic proximity between these concepts, yet, they cannot answer questions like: "To what extent the medical concept insulin captures the property of being a *pharmacological substance* or a *hormone* ?". In contrast, with the aid of proposed transformation technique (Figure 1), we can precisely answer such questions. The main advantage of having such form of post-hoc reasoning is that these interpretable representations might not only aid in generating explainable answers to the sensitive downstream medical tasks such as disease diagonosis [4], but also provide us with keen insights into the nature of state-of-the-art embedding models themselves. Motivated with these speculations, in this study, we consider the problem of improving the interpretablity of words embeddings learned over a particular text corpora.

Unlike numerous studies done on generating vector representations, literature on learning interpretable word embeddings is relatively scarce: [8]–[10]. In general, the key idea of these prior studies to improve the interpretability of vector representations is by inducing sparsity in the word vector dimensions [8]. Arguably, these studies made substantial advances, however, they still have a few inherent drawbacks. First, either these models cannot be learned over pre-trained word vectors available from the widely used embedding models such as Word2Vec [2]/GloVe [11] or they produce vectors with much higher dimensions. Second, these studies did not attempt to elucidate the particular conceptual notion (property) being carried within these individual dimensions.

To mitigate these aforementioned issues, in this study, we systematically formulate this problem of improving the intepretability of word embeddings. Basically, the core idea of the proposed model is to leverage upon the rich categorical/taxonomic knowledge present in the biomedical domain and learn a transformation matrix being sensitive to them. As the available categorical knowledge is manually curated and maintained by subject-matter-experts, our conjecture is that the interpretability of word embeddings in terms of these human-defined categories will reflect more proximity to the human level interpretations. Towards this end, we propose a novel framework that first infers the vector representation of

*Equal contribution

978-1-5386-9159-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDM.2018.00135

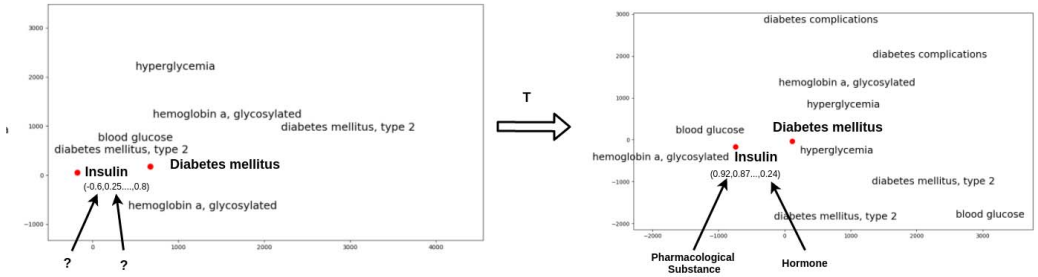

Fig. 1: The original word embedding space (left) and the transformed embedding space (right).

categorical concepts and then learns a transformation matrix that is able to transform the original word embeddings to a new space where these aforementioned categorical concepts act as their basis. Besides, the learning of transformation matrix is performed in such a way that the expressive features of original vectors are retained.

In this study our contributions can be summarized as:

1) We propose a novel framework for interpreting word embeddings, that is capable of transforming any pre-trained word embedding to a new space such that the hidden conceptual meaning of individual dimensions are revealed. To the best of our knowledge, we are among the first to study the interpretability of word embedding in the medical domain.
2) By leveraging upon the principles of dictionary learning and exploiting the categorical knowledge present in the biomedical domain, the proposed technique learns to infer the categorical representations at a granular level.

II. Related Work

Improving interpretability of word embeddings has been an active area of study over the past few years [8], [10], [12]. The initial study [12] proposed a non-negative matrix factorization based technique, namely, Non-Negative Sparse Embedding (NNSE) to learn the interpretable embeddings. While this study elucidated the importance of studying interpretability of word embeddings, yet, they were shown to suffer from memory and scale issues. To alleviate this, [13] proposed to learn interpretable embeddings in an online manner. In doing so, their key idea was to adopt a neural network approach to learn the word embeddings, and then employ an adaptive gradient descent to accelerate their convergence.

Building upon the ideas of aforementioned studies, [8] proposed a principled sparse coding technique to improve the interpretability of word vectors. Basically, they utilized sparse coding in a dictionary learning setting to obtain longer, sparser and overcomplete vectors. A potential drawback of this study is that it produces vectors of very high dimensions. More recently, another study [14] adopted l_1 regularization into their learning objective to induce sparsity and learned interpretable vectors. In general, the central notion behind these sparsity inducing techniques is that they aim to increase the sparseness in vectors, that then leads to better separability, thereby, improving the interpretability. While crucial insights were gained from these aforementioned studies, they still did not focus on explicating the precise conceptual/categorical meaning being carried within the individual dimensions. In this study, by relying upon the principles of categorical theory [15] and correspondingly exploiting the rich categorical knowledge present in the medical domain we attempt to study the interpretability of word embedding dimensions at a more granular level.

The work much akin to ours is a recent study done by [10]. In this study, the authors proposed to rotate the original vector dimensions in such a way that the rotated vectors are interpretable. While close in spirit, we differ from this study in two aspects. Firstly, the objectives are different. We aim to study the interpretability of words embedding in the medical domain by leveraging upon the categorical knowledge. Secondly, our problem is more difficult in a sense that the we aim to particularly illuminate the implicit conceptual notion remaining hidden within these individual dimensions.

III. Overview of Proposed Model

Recall that the input to our system is a set of pre-trained word vectors of medical concepts, and the goal is to learn a transformation matrix that projects the input embeddings to a new space wherein the transformed embeddings are both interpretable and retain their original expressive features.

To accomplish our first objective (interpretability), we focus on exploring the principles of category theory [15] and aim to interpret the embeddings in terms of these categories. Such categories in the biomedical domain refer to a broad subject themes that provide a consistent categorization of the medical concepts [16]. These categories in addition to possessing a conceptual meaning also have dictionary definitions associated with them. By taking advantage of this expert knowledge, we infer their categorical representations. These inferred categorical representations then further act as the basis for our new space. Once this new space is defined, we then learn a transformation matrix from the original embedding space to this new target space. This transformation matrix in particular allows us to achieve interpretability for the input embeddings in the transformed space.

Next, to achieve our second objective (i.e., retaining the expressive features present in the pre-trained vectors), a form

of orthogonal constraint is imposed on the learned transformation. Such form of imposition allows us to minimize the possible loss of information; thereby, aiding us to achieve our second objective of retaining the expressive information present in the pre-trained vectors. Further details on these are provided in Section IV-A and Section IV-B.

Last but not the least, we wish to highlight that one crucial advantage of adopting this transformation based technique is that it provides the proposed model an added flexibility of acting as a "plug-and-play" module for other downstream tasks. Because the proposed approach does not jeopardize the word embedding training process, it allows end-users the liberty of choosing their own method of generating word embeddings and then utilize the proposed model as a means of post-processing step to gain interpretability.

IV. Methodology

Our methodology section is divided into two sections. Section IV-A describes the technique to infer the categorical representations. Then, Section IV-B presents the details on how the transformation matrix is learned, and further discusses on how it induces the interpretability for word embeddings.

A. Inferring Categorical Embeddings

To infer the embeddings of categories, we leverage upon the dictionary definitions provided by the subject matter experts [16]. As an illustration, consider the definition of category "Disease or Syndrome" shown below:

Disease or Syndrome: *"A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder. Any specific disease or syndrome that is modified by such modifiers as acute, prolonged, etc. will also be assigned to this type. If an anatomic abnormality has a pathologic manifestation, then it will be given this type as well as a type from the Anatomical Abnormality hierarchy, e.g., Diabetic Cataract"*.

As these definitions are very precise, we leverage this expert knowledge and aim to infer the representation of "Disease or Syndrome". To do so, we first extract the medical concepts from their definitions and then use their already available word representations to infer their categorical meaning. Note that these medical concepts (underlined in the above example definition) are also present in our input pre-trained embeddings. Now, as the number of concepts contained in these categorical definitions is limited, this inevitably leads to a coarser estimation of their categorical meaning. To overcome this issue, we expand the set of associated medical concepts based on the external knowledge graph present in the bio-medical domain (the effectiveness of incorporating the neighbourhood set is validated in the experimental section). In this knowledge graph, the medical concepts are arranged in the form of a hierarchical tree (i.e., IS-A relationships). As such, the distance between the concepts in this tree denotes their semantic proximity. Building upon this premise, we assume that the concepts closer to each other in the hierarchy share greater information and thus the subtle cues obtained from the local neighborhood of concepts present in dictionary definitions might improve the overall categorical representation.

Formally, let $\mathbf{C} \in \mathbb{R}^{d \times m}$ denote the overall collection of categorical embeddings, d denote the embedding dimension and m denote the number of semantic categories. Now, to incorporate the above discussed local neighborhood information for concepts present in the dictionary definitions, a simple graph based scenario is considered. In this graph, nodes refer to the set of medical concepts and an edge is formed between concepts, if there is an hypernyms/hyponyms relationship between them. Let $\mathbf{V}_i = \{\mathbf{v}_{i1}, .., \mathbf{v}_{ij}\}$ denote the set of embedding vectors for medical concepts contained in the definition of i-th categorical concept, and $Neigh(\mathbf{v}_{ij})$ denote the corresponding set of local neighbours (siblings, parents and children) for the medical concept v_{ij}.

Our objective now is to infer the set of categorical embeddings $\hat{\mathbf{C}} = [\hat{\mathbf{c}}_i, ..., \hat{\mathbf{c}}_m]$ such that the categorical vectors are both close to the concepts present in their dictionary definitions and also to the local neighbours of the dictionary concepts. To achieve this, we propose the following loss function to infer their categorical representations:

$$L_c = \sum_{i=1}^{m} \left[\sum_{j=1}^{J} (||\hat{\mathbf{c}}_i - \mathbf{v}_{ij}||_2^2 + \sum_{k \in Neigh(\mathbf{v}_{ij})} \alpha ||\hat{\mathbf{c}}_i - \mathbf{v}_{ijk}||_2^2) \right] \tag{1}$$

where J denotes the number of dictionary concepts present in the particular category definition, and $\mathbf{v}_{ij}$, $\mathbf{v}_{ijk}$ represents the embeddings of dictionary concepts. The value of α is empirically set to 0.1 and is used to control the relative strengths between the concepts explicitly present in the dictionary definitions and their local neighbours. As it can be observed, the formulation is convex and its solution can be found by solving a system of linear equations. We minimize the categorical loss function and infer the categorical embeddings as follows:

$$\hat{\mathbf{C}} = \arg\min_{\hat{\mathbf{C}}} L_c \tag{2}$$

The entire set of categorical embeddings is denoted as $\hat{\mathbf{C}} = [\hat{\mathbf{c}}_1, ..., \hat{\mathbf{c}}_m]$, and the closed form solution for $\hat{\mathbf{c}}_i$ is shown below:

$$\hat{\mathbf{c}}_i = \frac{\sum_{j=1}^{J} (\mathbf{v}_{ij} + \alpha \sum_{k \in Neigh(\mathbf{v}_{ij})} \mathbf{v}_{ijk})}{J + \alpha \sum_{j=1}^{J} K_{ij}} \tag{3}$$

where K_{ij} represents the size of $Neigh(\mathbf{v}_{ij})$.

B. Learning Transformation

To be precise, we expect our transformation technique to meet the following two objective: *1) the implicit conceptual property within the individual dimensions should be enlightened* and *2) the transformation should be carried out in such a way that the resultant embeddings retain the information present in the original vectors.* To accomplish the first goal, the idea is to attain a target space (after performing transformation) with the basis as the semantics of inferred categorical

representations (refer Section IV-A). The corresponding value on individual dimension quantifies the amount of conceptual property being captured within these individual dimensions. Let $T : R^d \rightarrow R^m$ represent a linear transformation, and the transformed categorical embeddings are represented as $T(\hat{\mathbf{C}}) = [T(\hat{\mathbf{c}}_1), ..., T(\hat{\mathbf{c}}_m)]$. Since the transformed categorical embeddings act as a basis of the new space and these basis are also linearly independent unit vectors in the new space, an identity matrix could be used as a target for the transformed basis. To achieve this, we formulate the transformation as an optimization problem shown below:

$$\min_{\mathbf{W}} ||\mathbf{W}^T \cdot \hat{\mathbf{C}} - \mathbf{I}||_2^2 \tag{4}$$

Here the transformation matrix is denoted as $\mathbf{W}$ and $\mathbf{I}$ refers to an identity matrix. Note that this step acts as a soft regularization for linear independence, as in the real-word scenario, the distinct categorical embeddings may not be strictly independent of each other. In essence, this particular step of categorical basis conversion plays a vital role in inducing the interpretability in word vectors, and also allows us to explicitly define the meaning of the individual dimensions with their categorical types; thereby, enabling us to achieve our objective of performing dimension-wise interpretability.

Next, to meet our second objective of preserving the expressive features, we propose to regularize the transformation matrix by an orthogonal constraint. This is because of the peculiar property of orthogonal transformation to preserve the bilinear form i.e., Euclidean distance and Cosine in the latent space [10]. Since our transformation is from the original embedding space to an interpretable space, this may result in the change in number of dimensions; thereby, causing a possible information loss. To handle this, we adopt the principles of orthogonal transformation and mould that into our proposed optimization framework. This allows us to preserve the information particularly relevant to the categorical dimensions. The proposed orthogonal constraint is shown below:

$$\min_{\mathbf{W}} ||\mathbf{W}^T \cdot \mathbf{W} - \mathbf{I}||_2^2 \tag{5}$$

Now, since the focus of this study is to find a transformation matrix $\mathbf{W} \in R^{d \times m}$ that transforms the original pre-trained embeddings from d dimensional space to m dimensional space (that has inferred categorical embeddings as the basis), and the corresponding transformation matrix also attempts to preserve the information, the final objective to be minimized becomes the combination of these two objectives:

$$L_w = ||\mathbf{W}^T \cdot \hat{\mathbf{C}} - \mathbf{I}||_2^2 + \beta||\mathbf{W}^T \cdot \mathbf{W} - \mathbf{I}||_2^2 \tag{6}$$

Here β (empirically set to 0.2) controls the relative strengths of associations. To solve this, we take the gradient of our objective function (Equation 6) with respect to each of the model parameters and then adopt stochastic gradient descent to update our transformation matrix W:

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{\partial L_w}{\partial \mathbf{W}} \tag{7}$$

where η (empirically set to 0.001) is the learning rate for gradient descent. Overall, the fulfillment of two above discussed objectives allows us to achieve our goal of inducing the interpretability in vector representation and concurrently retain the original expressive features.

V. Experiments

The focus of this section is to demonstrate the efficacy of the proposed model in improving the interpretability of the pre-trained word embeddings. In doing so, we first need a set of word embeddings trained on a massive corpora. For this purpose, we choose MEDLINE[1] - the largest available bibliographic repository in the domain of biomedicine. At this time of writing, it contains more than 24 millions records (articles) primarily from the research area of life sciences and biomedicine. Every article in MEDLINE is tagged with a set of special keywords known as Medical Subject Headings (MeSH) terms. Because they are assigned by subject-matter-experts, they find their utility in a variety of biomedical tasks. Thus, we believe that the use of MeSH terms (and correspondingly release of interpretable MeSH embeddings[2]) will have immediate practical benefits to the community.

Based on the full-scale MEDLINE corpus (and correspondingly MeSH terms), we use CBOW [17] word embedding model to train our embeddings. Additionally, as means of an alternate baseline, we also train another prominent word embedding model, namely, GloVe [11] on the same MEDLINE corpora. As suggested by the previous studies [2], [11], the number of embedding dimension is set to 300. Also, note the total number of semantic types (m) available is 133 [16].

A. Interpretability

(1) Qualitative Assessment of Interpretability

To perform the qualitative assessment of our results, we borrow experimental settings from the interpretable word embeddings literature [8], [10]. Specifically, the idea in qualitative evaluation is that if a particular vector dimension is interpretable then the top ranking words (from the entire vocabulary) for that dimension should display a form of semantic coherence. To examine this, we select four examples of biomedical significance [6], [7]. The selected examples are the following: a) Diabetes mellitus, b) Migraine disorders, c) Alzheimer's disease and d) Insulin. For each of these examples, we examine their top participating dimension and then look up for the top words with highest value in the same dimension. Table I presents the results for both pre-trained word embedding models (both CBOW and Glove) and the proposed model. As it can be observed, the semantic grouping of words resulted by CBOW/Glove is more or less arbitrary. In contrast, the results obtained by our transformed embeddings yields a meaningful semantic coherence. As an illustration consider the case of "Diabetes mellitus". For the proposed model, it can be observed (refer Table I) that most of the terms

[1]https://www.nlm.nih.gov/pubs/factsheets/medline.html
[2]https://github.com/kishlayjha/InterpretableMedicalEmbeddings

TABLE I: Qualitative evaluation of the original and generated embeddings

Concepts	CBOW	Glove	Proposed
Diabetes mellitus	25-hydroxyvitamin d 2, 3-hydroxyacyl coa dehydrogenases, whiplash injuries, youth sports, abdominal fat	humans, xanthomatosis, cerebrotendinous, glycogen, yang deficiency	diabetes insipidus, diabetes complications, diet therapy, digestive system diseases
Indomethacin	acute kidney injury, acute disease, "administration, oral", "abortion, septic", acidosis	acetohexamide, "administration, intravenous", agglutination, albumins, "4-aminobenzoic acid"	endothelin-1, endothelins, endothelin-1, endotoxemia, "endothelin-converting enzymes"
Alzheimer Disease	ac133 antigen, acinar cells, ablation techniques, abducens nerve diseases, acinar cells	"active transport, cell nucleus", "acid sensing ion channels", "abducens nerve diseases", "acinar cells", "actins"	amyotrophic lateral sclerosis, amyloidosis, "amyloidosis, familial", amyloid neuropathies, "amyloid neuropathies, familial"
Insulin	alpha-msh, artemia, anabolic agents, antithyroid agents, appetite	appetite, acromegaly, adrenalectomy, anabolic agents, andropause	insulin antagonists, insulin-like growth factor binding protein 1, insulin-like growth factor binding protein 2, lactation, lactation disorders

in the group are closely related to the various aspects of "Diabetes" itself and the remaining few are related to the concept of "Disease" in general. In our transformed embeddings, we find the category name of these terms to be "Disease or Syndrome". Recall that as our transformation matrix is augmented with the categorical information, every dimension in the transformed vector is regularized by a particular categorical concept.

Another point we wish to highlight is the ability of the proposed model to answer question like: "To what extent a medical concept (e.g., Insulin) encodes the semantics of category *Pharmacological substance* or a *Hormone* within their dense dimensions". Note that the transformed embeddings have numerical values in their individual dimensions. These numerical values precisely help us in answering such aforementioned kind of questions. As an illustration, consider the case of "Insulin". In the medical domain, "Insulin" acts both as a pharmacological substance and a hormone. In our results, we obtained highest score for insulin in the category name - "pharmacological substance" and a relatively higher score in the category name - "hormone". From this result, one can speculate that the vector representations (generated by the state-of-the-art embedding models) of insulin captures the conceptual property of being a "pharmacological substance" more than that of a "hormone".

In essence, from the above discussed qualitative assessment it can be deduced that the proposed model is able to elucidate the meaning of individual dimensions and potentially shed insights into the notion of conceptual properties being captured by the state-of-the-art embedding models too. While informative, this form of qualitative assessment still does not inform us about the overall quality of the result set. To this end, a quantitative evaluation has to be performed.

(2) Quantitative Assessment of Interpretable Embeddings
In order to perform a quantitative assessment, we analyze our results on a task much akin to semantic categorization. In more detail, every medical concept present in our vocabulary belongs to a certain number semantic categories. For instance, the medical concept "Diabetes mellitus" belongs to the semantic category of "Disease or syndrome". In this manner, every concept present in the dictionary is assigned a semantic category from the range of one to five. We probe whether the dimension with highest score (i.e., semantic labels predicted by proposed model) match the true semantic labels or not. Table II reports the accuracy for our Top-K dimensions. Now, as the previous studies do not perform dimension-wise interpretability, a direct comparison with their approach cannot be performed. For the sake of comparison, we developed a baseline (i.e., Supervised) that uses all the explicit semantic labels to train a linear model (using pre-trained emebddings) and reported the results in Table II. As it can be observed, the proposed model (though unsupervised in nature) still maintains a reasonable performance as compared to the supervised model. Note that in our proposed model we do not use any explicit semantic labels. Now, in order to explore the effectiveness of incorporating the neighbour sets of medical concepts from the knowledge graph (refer Section IV-A), we evaluate the proposed model (with/without neighbourhood set) and report results. As it can be observed, the proposed technique of categorical inference significantly outperforms the baseline. We believe this is due to the ability of the proposed technique to obtain subtle cues from the informative neighbours of the dictionary concepts that ultimately improves the quality of categorical representation.

TABLE II: Quantitative evaluation of semantic categorization task

Baseline	Accuracy (K=5)	Accuracy (K=10)	Accuracy (K=15)
Supervised	0.732	0.857	0.925
Proposed model (without neighbours)	0.423	0.557	0.652
Proposed Model	0.522	0.683	0.762

In summary, from Section V-A, we can conclude that the proposed model has the capability to generate interpretable embeddings that have high proximity to the human intuition. While this accomplishes our core objective, we also aim to ensure that the information present in original pre-trained word vectors is retained in the transformed embeddings. To evaluate this, in Section V-B, we report and analyze our results on the biomedical concept similarity/relatedness tasks.

B. Expressive Performance

In this section, we inspect the expressive performance of our transformed embeddings as compared to the original vectors.

TABLE III: Absolute values of correlation of the five measures relative to human judgments - MeSH-1

Measure	Physician	Expert
CBOW	0.8174	0.7632
GLoVe	0.8057	0.7541
Proposed model	0.8015	0.74328

TABLE IV: Absolute values of correlation of the five measures relative to human judgments- MeSH-2

Measure	Human expert
CBOW	0.7677
GLoVe	0.7586
Proposed model	0.7789

(1) Evaluation Datasets

To examine the ability of transformed embeddings to retain original information, we choose biomedical concept similarity/relatedness task. The evaluation sets (i.e., MeSH-1 and MeSH-2) are borrowed from [18] and [19] respectively. Both datasets consist of 30 and 36 concept pairs that were manually rated by human experts indicating their semantic similarity.

(2) Results and Discussion

1) MeSH-1: Table III presents the Spearman (ρ) coefficient values obtained after applying the proposed model on the first dataset (MeSH-1). As it can be observed from the table, the proposed model performs on par with both CBOW and GloVE and achieves similar correlation as pre-trained embeddings with both physician's and experts judgments.

From the results, it can be inferred that the transformed embeddings retain the features of original vectors. We believe the reason for this lies in the orthogonality constraint imposed on the learned transformation. Because such form of imposition leverages the principles of orthogonal transformation (the has unique capability of preserving the bilinear form), the categorical related information loss is minimized.

2) MeSH-2: Table IV shows the correlation values obtained for the Spearman (ρ) coefficient for MeSH-2 dataset. Note that in this dataset, the proposed model obtains even higher correlation value as compared to the state-of-the word embedding models. Analyzing this result further, we believe that the reason for this lies in the capability of the proposed model to preserve the relevant information related to categorical dimensions in the transformed space, and correspondingly removing the unrelated information.

VI. Conclusion

In this study, we proposed a novel framework that induces the interpretability of word embeddings in the medical domain. Specifically, by leveraging upon the principles of category theory and rich categorical knowledge present in the biomedical domain, the model learns a transformation matrix that induces the interpretability of word embedding dimensions at a granular level. The transformation matrix in particular is learned in a such a way that any pre-trained input embeddings can be transformed to a new space where the produced embeddings reveal the conceptual meaning hidden within their individual dimensions and concurrently posses the expressive features present in the original pre-trained vectors.

Acknowledgment

This work was supported in part by the US National Science Foundation under grants NSF IIS-1218393 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[3] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang, "Wave2vec: Learning deep representations for biosignals," in *Data Mining (ICDM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1159–1164.

[4] X. Liu, D. Tosun, M. W. Weiner, N. Schuff, A. D. N. Initiative *et al.*, "Locally linear embedding (lle) for mri based alzheimer's disease classification," *Neuroimage*, vol. 83, pp. 148–157, 2013.

[5] D. L. Ngo, N. Yamamoto, V. A. Tran, N. G. Nguyen, D. Phan, F. R. Lumbanraja, M. Kubo, and K. Satou, "Application of word embedding to drug repositioning," *Journal of Biomedical Science and Engineering*, vol. 9, no. 01, p. 7, 2016.

[6] G. Xun, K. Jha, V. Gopalakrishnan, Y. Li, and A. Zhang, "Generating medical hypotheses based on evolutionary medical concepts," in *Data Mining (ICDM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 535–544.

[7] V. Gopalakrishnan, K. Jha, G. Xun, H. Q. Ngo, and A. Zhang, "Towards self-learning based hypotheses generation in biomedical text domain," *Bioinformatics*, 2017.

[8] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. Smith, "Sparse overcomplete word vector representations," *arXiv preprint arXiv:1506.02004*, 2015.

[9] S. Rothe, S. Ebert, and H. Schütze, "Ultradense word embeddings by orthogonal transformation," *arXiv preprint arXiv:1602.07572*, 2016.

[10] S. Park, J. Bak, and A. Oh, "Rotated word vector representations and their interpretability," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 401–411.

[11] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[12] B. Murphy, P. Talukdar, and T. Mitchell, "Learning effective and interpretable semantic models using non-negative sparse embedding," *Proceedings of COLING 2012*, pp. 1933–1950, 2012.

[13] H. Luo, Z. Liu, H. Luan, and M. Sun, "Online learning of interpretable word embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1687–1692.

[14] F. Sun, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Sparse word embeddings using l1 regularized online learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2915–2921.

[15] G. Murphy, *The big book of concepts*. MIT press, 2004.

[16] A. T. McCray, A. Burgun, and O. Bodenreider, "Aggregating umls semantic types for reducing conceptual complexity," *Studies in health technology and informatics*, vol. 84, no. 0 1, p. 216, 2001.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[18] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of biomedical informatics*, vol. 40, no. 3, pp. 288–299, 2007.

[19] A. Hliaoutakis, "Semantic similarity measures in mesh ontology and their application to information retrieval on medline," *Master's thesis*, 2005.