This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

Visual Exploration of Neural Document Embedding in Information Retrieval: Semantics and Feature Selection

Xiaonan Ji, Han-Wei Shen, *Member, IEEE,* Alan Ritter, Raghu Machiraju, and Po-Yin Yen

Abstract—Neural embeddings are widely used in language modeling and feature generation with superior computational power. Particularly, neural document embedding - converting texts of variable-length to semantic vector representations - has shown to benefit widespread downstream applications, e.g., information retrieval (IR). However, the black-box nature makes it difficult to understand how the semantics are encoded and employed. We propose visual exploration of neural document embedding to gain insights into the underlying embedding space, and promote the utilization in prevalent IR applications. In this study, we take an IR application-driven view, which is further motivated by biomedical IR in healthcare decision-making, and collaborate with domain experts to design and develop a visual analytics system. This system visualizes neural document embeddings as a configurable document map and enables guidance and reasoning; facilitates to explore the neural embedding space and identify salient neural dimensions (semantic features) per task and domain interest; and supports advisable feature selection (semantic analysis) along with instant visual feedback to promote IR performance. We demonstrate the usefulness and effectiveness of this system and present inspiring findings in use cases. This work will help designers/developers of downstream applications gain insights and confidence in neural document embedding, and exploit that to achieve more favorable performance in application domains.

Index Terms—Neural document embedding, information retrieval, semantic analysis, feature selection.

1 Introduction

DOCUMENT representation is an important topic in text analytics and language modeling, aiming to encode essential features and underlying meanings of documents in a structured and machine understandable manner. Neural document embedding [1]–[3], as a successful extension to neural word embedding, has shown to leverage superior computational power of neural networks and generate effective document representations in concise feature vectors. In other words, neural embeddings not only mitigate the curse of dimensionality, but also demonstrate to capture text hidden semantics and outperform lexical and conventional embedding methods. The resulting semantic representations substantially benefit downstream applications [4], such as information retrieval (IR), sentiment analysis, sentence modeling, machine translation, etc. In particular, IR typically involves text similarity, clustering, or classification, and aims to identify relevant documents for an information need. It relies on effective document representations to capture document meanings and characterize document relevancy. The remarkable performance of neural embedding empowers IR to cope with the growing volume of information resources and fulfill critical needs.

Despite the superior performance in IR and other text analytics applications, neural document embedding is usually used as a black-box, and it is difficult to understand how the performance is achieved or how to tune the performance in different conditions. Many studies attempt to evaluate neural embedding models, but they generally rely on trial-and-error or benchmarks with limited characteristics [5]. Due to the presence of boundless analytic facets, noise, and redundancies in neural embedding, it is challenging to make thorough interpretations even with sophisticated mathematical tools. Thus, for prevalent IR applications, it remains unclear how underlying document meanings (e.g., semantics) are encoded in the hidden states (e.g., features or dimensions) in the embedding space, and how that contributes to IR performance, such as document semantics, similarity, classification, or clustering (of relevant documents). Moreover, we are motivated by IR applications in the biomedical and clinical domain, where advanced IR is needed to accelerate system review (SR) production in healthcare. In a real-world and critical SR, relevant biomedical documents (e.g., published studies or clinical trials) embracing high-quality research findings are retrieved to guide patient care and inform clinical decisions, thus supporting Evidence-based Practice (EBP). While the IR applications can benefit from advantageous neural document embedding, it remains unclear how semantics of domain and task interest, such as biomedical concepts and topics, are encoded and built-up in neural embedding.

Therefore, we use visual analytics to explore neural document embedding with two main goals: (1) understand the

- *X. Ji was with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, and is with the Institute for Informatics, Washington University School of Medicine, St. Louis, MO 63108. E-mail: ji.62@osu.edu.*
- *H.-W. Shen, A. Ritter, and R. Machiraju are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210. E-mail: {shen.94, ritter.1492, machiraju.1}@osu.edu.*
- *P.-Y. Yen is with the Institute for Informatics, Washington University School of Medicine, St. Louis, MO 63108, and Goldfarb School of Nursing, BJC Healthcare, St. Louis, MO 63108. E-mail: yenp@wustl.edu.*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

underlying mechanism, especially how semantics are encoded in the neural dimensions (e.g., hidden semantic features), and (2) with the interpretation, apply proper configurations to enhance IR performance via feature selection (semantic analysis) per domain or task interest. To achieve aforementioned goals, we take an application-driven view and collaborate with domain experts who cultivate (i.e., design and develop) IR applications in healthcare. We then formulate a series of visual analytics tasks and iteratively develop a visual analytics system. Our work is also inspired by: 1) visual analytics of neural models to elucidate the relations among neuron behaviors, learned features or representations, and the associated performance, and 2) visual analytics of document corpus to reveal interpretable properties and patterns of document features, representations, and relationships, thus promoting the exploration.

Among the different types of neural document embedding models, we use the Paragraph Vector (PV) model [1], which is a general-purpose model with demonstrated effectiveness in IR. Our study design is generalizable to document embeddings produced by other models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In summary, our work contributes to:

- Enable exploration and interpretation of neural document embedding, especially with respect to the encoded semantics of domain and task interest.
- Facilitate semantic feature selection in the embedding space to promote performance in IR applications.
- Validate a useful and effective visual analytics system, with use cases to enhance human interpretation and exploitation of neural document embedding.
- Apply visual text analytics in pressing real-world issues, i.e., biomedical IR and SR/EBP in healthcare.

2 BACKGROUND

2.1 Neural Document Embedding

Neural embedding leverages the advantageous learning power of neural networks and is widely used in language modeling and feature learning, where texts are mapped to vectors of real numbers. It conceptually involves a mathematical embedding that converts texts from an original feature space in high dimensions to a vector space defined in low dimensions. Neural embedding also intends to resolve a critical issue in language modeling: to capture underlying meanings and hidden semantics (features), and encode them into machine understandable representations. Under this notion, word embedding (e.g., word2vec) was approached first, and recent embedding models [1], [6] were studied for variable-length text, e.g., phrase, sentence, paragraph, and document. These neural models demonstrate superior performance compared to lexical or conventional methods, such as PCA, SVD, LDA, etc. They also show to augment the performance of many machine learning models, such as SVM, logistic regression, k-means, t-SNE, etc.

Existing neural document embedding models can be categorized based on their architectures, generally including CNNs, RNNs and their variants (e.g., LSTM and GRU) [3], [6], and the well-known Paragraph Vector (PV) model [1], [2]. Besides, neural models are trained with different tasks to capture semantics in a text corpus. Typically, deep neural models are better trained by supervised tasks; shallow models (e.g., PV) can also be well trained by unsupervised tasks [5]. Because of the high generalizability and competitive performance of PV in document embedding, we use it to demonstrate our visual analytics approaches.

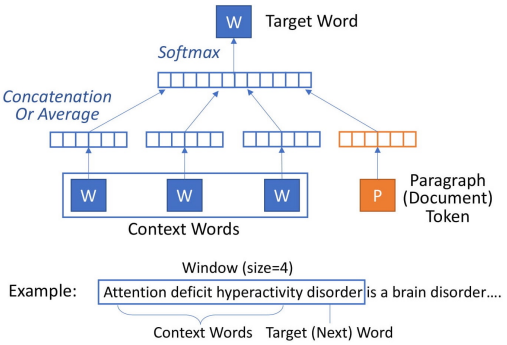


Fig. 1. An illustration of the Paragraph Vector model with distributed memory (PV-DM) [1].

2.2 The Paragraph Vector Model

The PV model [1], [2] is a 2-layer neural network inspired by word2vec. Paragraph refers to any length of text, ranging from phrase, sentence, paragraph, to document. It learns vector representations in an unsupervised way, leveraging distributed semantics conveyed by the context information, with an intuition that terms in similar contexts are semantically similar. It shows to produce competitive performance in IR especially for long text, e.g., document.

In the PV model, paragraph (document) representations are trained to predict words in the paragraphs. There are 2 types of PV models: (1) the distributed memory model (PV-DM), and (2) the distributed bag-of-words model (PV-DBOW). We use PV-DM in this work as it has demonstrated better performance. PV-DM (Figure 1) concatenates or averages a paragraph vector with a set of context word vectors to predict a target word in the context. A context is fixed-length and sampled from a sliding window over the paragraph. The paragraph vector is shared among all contexts and repeatedly updated towards the objective, thus it is considered a memory of all contexts or a topic of the paragraph. The PV-DM model has two key advantages: (1) it inherits the semantics from word vectors, and (2) it handles word orders and captures the semantics of local contexts.

We used the Python library *gensim* to implement PV-DM. For training instances, we took a window size of 8 (5-12 is commonly used [1]). For the embedding size or dimensionality, we took 200 (100-1000 is commonly used), and the PV model is less sensitive to the dimensionality [2].

3 RELATED WORK

3.1 Visual Analytics of Neural Networks

Our work is inspired by previous work in visual analytics of neural networks with respect to their intermediate or final results. CNNs in pattern recognition are first studied to gain insights into the behaviors of neurons and layers, and understand the superior performances and potential improvements. Zeiler et al. [7] analyzed layers' contributions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

to predictions, Simonyan et al. [8] visualized gradients and class saliency maps, Liu et al. [9] explored the multi-facets of neurons and learned features, Rauber et al. [10] analyzed learned representations and the contributing neurons, etc. Meanwhile, RNNs (LSTM and GRU) are widely exploited in NLP for sequence modeling. Related work has focused on analyzing activities or behaviors of hidden states, exploring captured linguistic aspects, and understanding the semantic properties. Karpathy et al. [11] and Li et al. [12] visualized cell activations (values) and revealed their contributions to different linguistic aspects or semantic meanings, Cirik et al. [13] analyzed internal states' behaviors in curriculum learning, Strobelt et al. [14] examined hidden state dynamics and the associated semantics, etc.

Visual analytics of neural embedding models has attracted increasing interest in recent years. Kiros et al. [15] visualized and qualitatively evaluated sentence representations by GRU, Smilkov et al. [16] developed an embedding projector to understand neural embeddings and their performances, Palangi et al. [17] visualized activation behaviors and resulting embeddings by LSTM and disclosed the encoding of keywords and topics, Lin et al. [18] revealed the encoding of important sentence components by LSTM, Dai et al. [3] visualized document representations by PV and evaluated clustered benchmark documents, Liu et al. [19] visualized word2vec's semantic analogies, etc.

Additional recent work has addressed both the interpretability and interactivity of neural models. Strobelt et al. [20] associated seq2seq model decisions and latent vectors with representative examples (neighborhoods) in datasets, and allowed users to improve the performance via customization. Kwon et al. [21] improved neural models applied to the healthcare domain, where medical codes and their contributions (attentions) to medical predictions were exposed to users for examination and improvement. Kahng et al. [22] addressed models in the industry, and presented a generalizable visual analytics system, where different analytical angles were enabled by flexible subset selection.

In summary, visual analytics of neural networks has focused on the relations among neurons, learned features or representations, and performance. Furthermore, in NLP, it is anticipated to understand semantic properties and neurons' contributions. Neural document embeddings are the final neuron states of the hidden layer, and it is expected to explain and exploit the encoded semantics and promote the interpretability and interactivity in applications.

3.2 Visual Analytics of Text Corpus

Our work is also related to visual analytics of text corpus to interpret and explore text features, representations, and relationships. Related work summarized text corpus and revealed document properties and patterns [23], using various types of visualizations for different purposes. Graph-based network [24], [25] and distance-based map [23], [26], [27] are used to layout documents in a 2D space with document relations encoded by the compelling spatial channel. For 2D embedding, t-SNE shows to retain important data structures (e.g., clusters) and is considered a preferred way to visualize document maps compared to some conventional methods (e.g., MDS). Furthermore, document clusters with content topics are considered the most interesting properties. Related work has enabled interactive explorations of clusters and topics. Cao et al. [24] allowed users to explore the multifaceted relations among documents and clusters. With a document map, Heimerl et al. [23] used focus+context to characterize groups of documents with keywords and topics, Kim et al. [26] presented an interactive lens interface with topic modelling on the fly, Ruppert et al. [28] allowed users to interactively create and validate clusters while exploring a corpus, etc.

In summary, visual document distribution, clusters, and topics are used to assist in understanding and exploring a corpus, as well as the underlying representations and features. We extend this idea and visualize a configurable document map, with keyword and topic synthesis for document clusters (or groups) in the neural embedding space.

4 DESIGN PROCESS

4.1 Domain Problem Characterization

We are motivated by prevalent and significant IR applications in the biomedical domain to support SR and ultimately promote EBP. EBP utilizes the best available research evidence to guide clinical practice (e.g., a treatment or procedure suitable for a patient). SR provides the highest quality of evidence for EBP by utilizing systematic approaches to identify, appraise, and summarize relevant biomedical documents (e.g., published studies and clinical trials). Healthcare providers and researchers use the findings (best evidence) from SR to implement EBP.

To characterize the domain problem and design our system, we worked closely with domain experts (E1 and E2) in biomedical informatics with both clinical and computational experiences as well as human-computer interaction knowledge. Specifically, E1 has a clinical background with extensive SR experiences in the biomedical domain. E2 has a computer science background, focusing on NLP in electronic health records. Both experts have cultivated (i.e., designed and developed) IR applications, especially for SR production. We held multiple design sessions to discuss IR tasks related to SR, and difficulties and advantages of using neural document embedding. Basically, an SR involves intensive IR tasks to identify relevant biomedical documents from a target corpus, which is typically obtained via exhaustive search on biomedical databases and consists of hundreds/thousands of documents (comprehensiveness). However, only a small percentage of these documents (accuracy) are considered relevant [29]. For example, a corpus contains documents with diverse disease subtypes, patient populations (P), interventions (I), comparisons (C), and outcomes (O). With specific PICO formulated in SR, the involved IR aims to identify documents relevant to the PICO. Due to the large-volume, heterogeneous, and noisy nature of biomedical documents, IR for SR suffers from limited effectiveness and efficiency, leaving critical issues unsolved.

There have been automated methods to assist in the identification of relevant documents, and the performance is evaluated on benchmark datasets, where documents that are relevant to a research question or patient problem (e.g., PICO in SR) are manually identified and labeled by at least

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

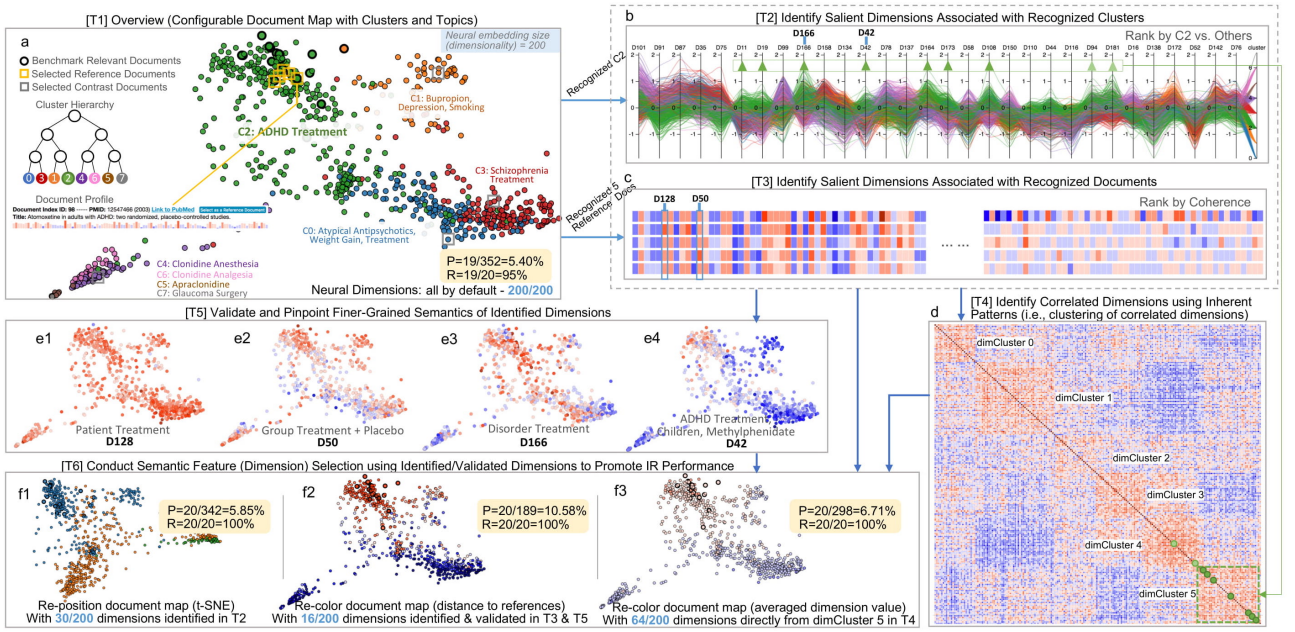


Fig. 2. A schematic overview of the visual analytics tasks and components. Details are provided in Sections 4 and 6.1.

two human experts (thus these documents are considered gold-standards or *benchmark relevant documents*) [29]. Prior to document classification, clustering, or ranking in downstream, documents are converted into representations to encode essential features and underlying meanings. While conventional feature engineering was commonly used, the experts agreed on the importance of leveraging advantageous neural embedding to promote document representations to a concise and semantic level. However, concerns remain for the interpretability and utilization (interactivity). Despite the good performances in many cases, domain experts expressed that *it was difficult to have complete confidence without knowing the semantics behind the embedding, especially how biomedical concepts for an IR task were captured*. Also, as an embedding space usually comes with diverse and noisy information, the experts found *their hands were tied when trying to recognize task favorable features, nor to make use of them to promote task-specific IR performance*.

With a user-centered design, the interests of domain experts can be summarized as to: (1) understand (assess) the performance of neural document embedding, (2) understand how the performance is achieved, i.e., via the encoded semantics, and (3) conduct semantic feature selection to promote performance toward an IR goal. We iterated through several prototypes and incorporated domain experts' feedback, such as to leverage interpretable information and patterns (e.g., keywords, topics, and clusters) and bridge an understanding of the embedding and hidden semantics; enable a guided top-down process to explore and identify salient dimensions, instead of trial-and-error; enable a bottom-up process to validate and pinpoint finer-grained semantics encoded in identified dimensions.

4.2 Analytics Tasks

We describe the visual analytics tasks (TI-T6) distilled from our design sessions. The tasks guided the development of our visual analytics system through iterative refinements. Figure 2 shows a schematic overview of the tasks.

T1: Obtain a basic understanding of neural document embedding by assessing the qualitative performance in IR. The rationale is to visualize a document corpus in the embedding (feature) space with interpretable clusters and topics, enabling a rapid understanding of the performance, i.e., clustering of similar documents. For multiple granularities, hierarchical clustering can be used to obtain an adjustable clustering level with corresponding topic synthesis. Such a mental map also enables a guided exploration.

T2: Identify salient neural dimensions (i.e., hidden semantic features) associated with recognized clusters and topics. For a guided and top-down exploration, clusters and topics can indicate (probe) high-level semantics across a corpus and allow users to determine the analytic facets per domain and task interest. This design rationale is inspired by NLP probing tasks [30] and aligns with [14], [19], [20], [22] such that abstract semantics and different analytic facets are made tangible by instances or clusters of certain properties. As dimension behaviors can be characterized by values across a corpus, we visualize patterns associated with clusters for users to identify salient dimensions.

T3: Identify salient neural dimensions associated with recognized documents. As another guided and top-down exploration, documents recognized with prior or external knowledge can also indicate domain and task interest and probe more particular semantics. Similar to T2, we visualize dimension behaviors and patterns associated with recognized documents for users to identify salient dimensions. T2 and T3 can be sequential or parallel actions.

T4: Identify correlated neural dimensions using inherent patterns. While T2 and T3 allow users to inject "semi-supervision" in the form of specifying clusters or documents of interest to identify salient dimensions, T4 leverages inherent dimension correlation in an "unsupervised" way to identify highly correlated dimensions. We visualize dimension correlation for users to explore the inherent patterns (independent action) and expand the identified dimensions with correlated ones (follow-up to T2 or T3).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

T5: Validate and pinpoint finer-grained semantics encoded in identified neural dimensions. Along with T2-T4, it is necessary to investigate and validate an identified dimension before applying it to semantic feature (dimension) selection. As a bottom-up action for reasoning, we allow users to specify a dimension, examine its value distribution across the corpus, and obtain topic synthesis from responsive (representative) documents to pinpoint the semantics.

T6: Conduct feature selection with identified or validated neural dimensions to promote IR performance. With salient or validated dimensions from T2-T5, users can exploit them in feature selection. The rationale includes not only reducing noise but also specifying domain and task interest for IR. With a subset of dimensions forming a new feature space, users can re-visit document relationships with an adapted document map, where the refined visual encodings might support an expedited identification of relevant documents for IR. Also, users can have instant visual feedback of the IR performance for iterative improvement.

In summary, the tasks were generated through a user-centered design: T1 provides an overview; T2 and T3 allow users to inject domain and task specifications to identify salient dimensions (features); T4 allows users to explore inherent patterns and expand salient dimensions; T5 provides validation and reasoning for T2-T4; and T6 leverages the knowledge from T1-T5 to promote IR performance.

4.3 System Overview

We present an overview of the visual analytics components (Comp1-Comp4) to support the analytics tasks. Mappings between tasks and components are indicated in brackets.

Comp1: Visualization of document embeddings. Visualize a configurable document map based on neural document embeddings. The document map can adapt to adjustable clustering (with dynamic topic synthesis) and a customizable embedding (feature) space. **[T1, T5, T6]**

Comp2: Global and cluster-level dimension behavior. Visualize dimension values across the corpus, and allow users to examine/rank dimensions based on global or cluster-level properties and identify salient dimensions. **[T2]**

Comp3: Document-level dimension behavior. Visualize dimension values across a set of documents, and allow users to examine/rank dimensions based on document-level properties and identify salient dimensions. **[T3]**

Comp4: Visualization of dimension correlation. Visualize dimension correlation matrix and reveal clusters of correlated dimensions, and allow users to explore the inherent patterns and expand identified dimensions with highly correlated ones. **[T4]**

5 VISUAL ANALYTICS COMPONENTS

5.1 Visualization of Document Embeddings

We present a visualization of document corpus to reveal document relationships and patterns in the neural embedding (feature) space. Clusters and topics are also imposed to promote interpretation and guide exploration. When adjusting the feature space, the map can adapt to reflect relationships and patterns in the new space, providing reasonings about the underlying features. Similar ideas to leverage human control (domain knowledge and task preference) for an improved map and clusters can be found in [31]. In IR, such a visualization can reveal clusters of similar documents and facilitate identifying relevant documents sharing similar features of domain and task interest.

5.1.1 Configurable Document Map

With n-dimensional (n=200) document embeddings produced by the PV model, we use t-SNE to reduce the dimensionality to two so that document points can be placed by their 2D coordinates and a 2D document map is visualized. Figure 3(a) shows a document map on the *attention deficit hyperactivity disorder (ADHD)* corpus from the *Drug Effectiveness Review Project (DERP)* [29]. The map is colored by document clusters in the neural embedding space.

The document map is configurable in terms of the positioning and coloring of document points. While all 200 neural dimensions (features) are applied to t-SNE by default, users can select any subsets of dimensions to re-generate the document map and re-position the documents for the specified feature space. When calling t-SNE, users can configure parameters such as the perplexity and learning rate, with Barnes-Hut approximation used for better efficiency. For coloring, documents are colored by their clusters which are configurable. The coloring can also adapt to other properties in a specified feature space. Therefore, the document map will support an exploration towards the IR performance (e.g., positioning/coloring of relevant documents) and the underlying neural dimensions (e.g., feature selection). Details will be provided in Section 5.1.2-5.1.4.

5.1.2 Hierarchical Clustering

Users can invoke and apply hierarchical clustering to neural document embeddings (with all or selected dimensions) or the 2D embeddings produced by t-SNE. As a bottom-up method, the agglomerative hierarchical clustering generates a dendrogram with multilevel clustering results. Figure 3(c) and (e) illustrate the selection of a clustering level (e.g., level=4), and Figure 3(a) shows the resulting clusters, where documents are colored by the clusters (e.g., 8 clusters). Users can adjust the granularity via the clustering level based on their need for coarser or finer-grained clusters. Alternatively, users can also use k-means clustering but must pre-specify the number of clusters (k).

Due to the approximation nature of t-SNE, documents from the same cluster, as established in the neural embedding space, are not necessarily placed close together in the 2D map, e.g., blur cluster boundary or visual distortion. To reinforce cluster patterns in the neural embedding space, users can aggregate documents towards their cluster center ($center_{cluster} - center_{global}$) with a multiplier α controlling the magnitude. Figure 3(f) shows 2D maps with different aggregation magnitudes (α). This simple method is not to alter t-SNE results, but rather to reinforce the clusters produced by neural embedding, while keeping the relative positions of clusters and intra-cluster documents by t-SNE. Related work to steer or augment a dimensionality reduction method, such as using cluster labels, constraints, or modified distance functions to preserve clusters or obtain better clusters, can be found in [32]–[35].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

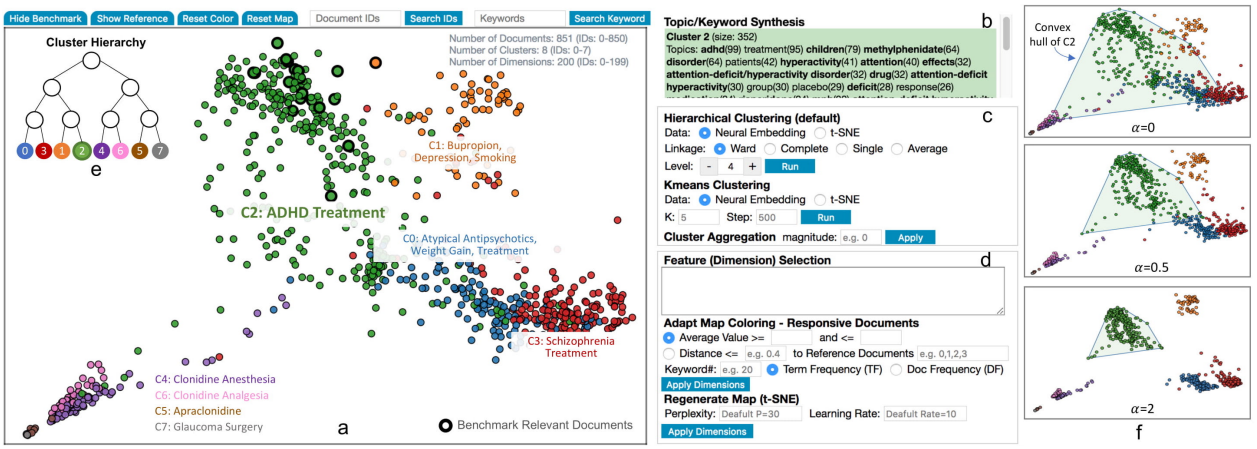


Fig. 3. Visualization of neural document embeddings on ADHD corpus. (a) Configurable document map, (b) Dynamic topic/keyword synthesis, (c) Document clustering, (d) Feature selection, (e) Dendrogram of multi-level clusters, and (f) Reinforced clusters in the embedding space.

5.1.3 Topic Synthesis

As shown in Figure 3(b), topics of clusters on a user selected level are dynamically synthesized from document keywords. Document keywords are generated by a RAKE-based method in prior text processing. Specifically, we re-implemented the RAKE (*Rapid Automatic Keyword Extraction*) algorithm [36] with parameter tuning and retrofits of noun phrases regarding biomedical documents. For each document, the method chunks texts into phrases and then scores and ranks the phrases to reflect their content importance. Therefore, in preliminary text analytics, noun phrases with high ranks in a document are considered the document's keywords; and in interactive visual analytics, keywords with high document frequencies in a cluster are used to form the cluster's topics. The RAKE-based topic synthesis is highly efficient and can generate results comparable to or even more informative than conventional topic modeling methods. For example, for the ADHD corpus [29], we list topics (top 10 keywords) of the cluster containing the most benchmark relevant documents, which have been recognized to address *ADHD treatments (e.g., drugs) and the efficacy on children/adults*. Compared to LDA and NMF, the RAKE-based method can generate more relevant and informative topics, as below highlighted in *italic*.

- LDA: risperidone, clozapine, olanzapine, *adhd*, placebo, *child*, symptom, haloperidol, *methylphenidate*, disorder.
- NMF: disorder, hyperactivity, *methylphenidate*, *child*, placebo, result, *adhd*, deficit, response.
- The RAKE-based method: *adhd*, *treatment*, *children*, *methylphenidate*, patient, hyperactivity, attention, *effect*, *attention deficit hyperactivity disorder*, *drug*.

Another topic synthesis work in public health was performed by Zhang et al. [37]. Besides topic synthesis for clusters, this method is also extended to any groups of documents, such as the responsive (representative) documents which will be described in Section 5.1.4. In this relation-seeking scenario, topics and keywords bridge the interpretation of hidden semantics, thus allowing users to infer (probe) semantics implied by a group of documents.

5.1.4 Feature Selection

By default, a document map is positioned by t-SNE and colored by hierarchical clustering with all dimensions (features) applied. Users can conduct feature selection to specify a subset of neural dimensions, and the visual encodings in the map will adapt to reflect properties and patterns in the new feature space (subspace). With a feature selection box, users can specify any dimensions or directly feed a set of identified dimensions, which will be described in Sections 5.2 - 5.4. Once dimensions are selected, there are 3 ways to adapt the document map, as shown in Figure 4.

(1) Re-color the default document map (positioned by t-SNE using all dimensions) so that each document x is colored by the averaged dimension value over the selected dimensions, e.g., $(1/|sel|) * \sum_{i \in sel} v_{x,i}$, where $v_{x,i}$ is the ith dimension value in x's embedding vector, and sel is a set of selected dimensions. (2) When some reference documents (e.g., benchmarks) are available, re-color the default document map based on a document x's similarity (distance) to the references, considering the subspace with selected dimensions, e.g., $cosine(v_{x,\{sel\}}, (1/|ref|) * \sum_{r \in ref} v_{r,\{sel\}})$, where $v_{x,\{sel\}}$ is a sub-vector of x's embedding vector with selected dimensions, ref is a set of reference documents, and *cosine similarity* is used. Both (1) and (2) use a heatmap coloring scheme, which is also referred to as a dot map or stress map [38]. Users can determine the *value or distance threshold* to retrieve *responsive (representative) documents* meeting certain criteria. And topic synthesis will be triggered to bridge the semantics carried by the responsive documents and help infer the semantics encoded in the selected dimensions. (3) Apply the selected dimensions to t-SNE, which will re-generate the document map encoding refined document relationships and clusters in the new feature space, with both the coloring and positioning adapted.

With (1), users can also specify a dimension to explore its value distribution across the document map colored by the dimension's value, and the responsive (representative) documents used for topic synthesis can be obtained by a value threshold or a quartile. This allows users to validate and pinpoint a dimension's semantics before utilizing it.

5.2 Global and Cluster-Level Dimension Behavior

The behavior of dimensions can be characterized by their values across a corpus. With a recognized cluster / topic (semantics) of interest, users can examine associated patterns, rank dimensions, and identify salient dimensions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

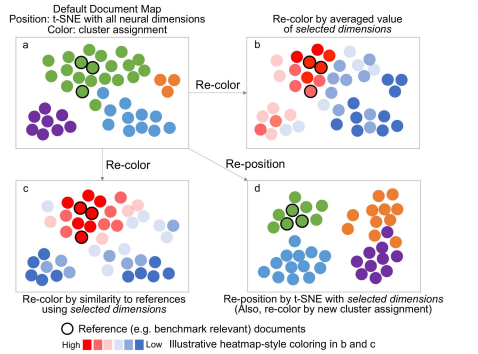


Fig. 4. Adapted document map by feature selection.

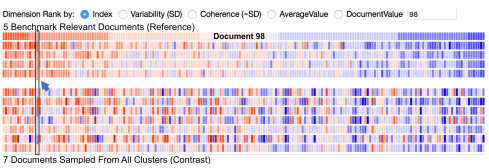


Fig. 6. Document-level dimension behavior. Dimensions are ranked by values in one benchmark relevant document (e.g., document 98).

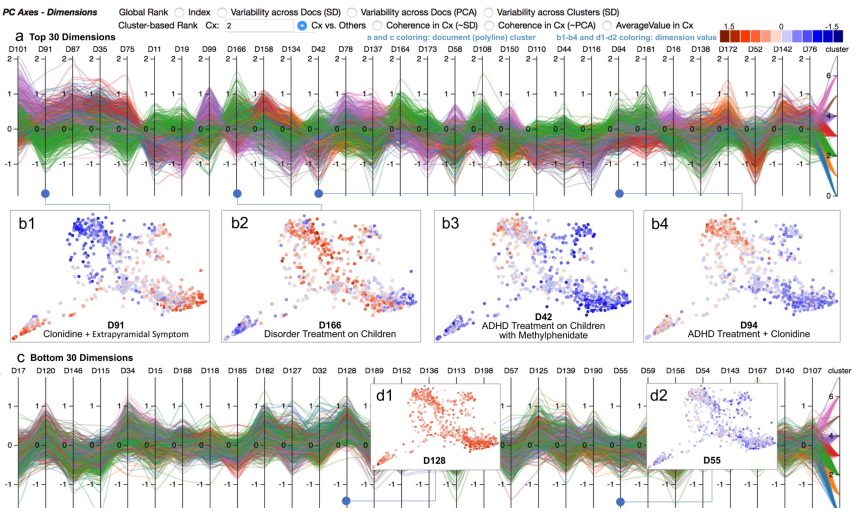


Fig. 5. Cluster-level dimension behavior, cluster C2 (green) is recognized to be of interest. (a) Top 30 dimensions ranked by *C2 vs. Others*. (c) Bottom 30 dimensions ranked by *C2 vs. Others*. (b1)-(b4), (d1)-(d2) Adapted document map with probed semantics for a selected dimension; a document point is colored by its value in the selected dimension.

5.2.1 Parallel Coordinates

As shown in Figure 5, we use parallel coordinates (PCs) to visualize dimension behaviors across a corpus (e.g., all documents) by representing each dimension as a vertical axis (coordinate) and each document as a data item (polyline). Polylines are colored by documents' cluster assignments and positioned by documents' dimension values. Heatmaps were used in related work but are limited to a smaller number of items. PCs can accommodate a larger number of items and aggregate polylines to form attentive spatial patterns. For example, documents with similar dimension values may form overlapped (reinforced) regions.

The inherent order of the dimensions resulted from neural embedding is meaningless, and we only use it for indexing, i.e., D0 - D199. To identify dimensions in a guided manner, we provide multiple options to rank dimensions based on properties of interest (described in Section 5.2.2). For better scalability and resolution, users can specify a range of dimensions from the ranked list to be displayed, for example, the top 30 or bottom 30 dimensions as shown in Figure 5 (a)(c). Users can also specify the documents to be displayed, which reduces clutter and allows users to concentrate on certain documents. They can brush vertical axes to select documents by dimension values or use the rightmost axis to select by clusters.

In our earlier prototypes, some alternative designs included: (1) Use of error bars to encode dimension variability (with max, min, mean, and quartiles). This has limited capability to encode values in particular documents/clusters. (2) Use of PCs where documents are vertical axes and dimensions are polylines. This has limited scalability when all or many documents are involved, and it is difficult for users to set criteria to rank documents and gain patterns.

5.2.2 Dimension Rank

We rank dimensions based on global or cluster-based behaviors when a cluster of interest is recognized. Intuitively, with a limited embedding size, high-quality embeddings should differentiate documents across a corpus and effective dimensions should account for the variability. For a recognized cluster, effective dimensions should account for intra-cluster coherence and inter-cluster difference.

The global ranks are for users to obtain an overview of dimension behaviors. To capture the variability of a dimension's value across all documents, we calculate the standard deviation (SD) for each dimension. A dimension with a larger SD will have a higher rank, representing a higher variability. As an alternative approach and justification, we also implement PCA across all documents, and rank dimensions based on their contributions to the top principal components (the largest possible variability). We found these two approaches had a high agreement (e.g., 73.33%) especially on the top-ranked dimensions (e.g., top 30). In addition, we rank dimensions by SD across all cluster means, to better capture the variation across clusters.

The cluster-based ranks are for users to identify salient dimensions associated with a recognized cluster/topic of interest. For example, users may be interested in finding the top dimensions best differentiating a particular cluster from all others. Therefore, once users specify a cluster Cx, we extend the idea of SD and use cluster means to calculate how other clusters are deviated from Cx. The deviation is formulated as a weighted sum of squared distances, giving larger clusters a larger weight. We call this rank *Cx vs. Others*, and the deviation score for a dimension d is:

$$dev(c_x, d) = \sum_{c_i \in C, c_i \neq c_x} size(c_i) \cdot (mean(c_i)_d - mean(c_x)_d)^2$$

Figure 5(a) and (c) illustrate the top 30 and bottom 30 dimensions ranked by *C2 vs. Others*. As the polylines representing documents are colored by their clusters, C2 documents are in green (same with Figure 3). Users can observe that the top 30 dimensions tend to form overlapped green regions which are (partly) isolated from polylines in other colors, reflecting the discrimination by these dimensions. In contrast, the bottom 30 dimensions tend to have highly mixed polylines, indicating worse discrimination. Besides, we provide SD and PCA-based ranks with respect to the documents in Cx. For intra-cluster coherence, the inverse

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

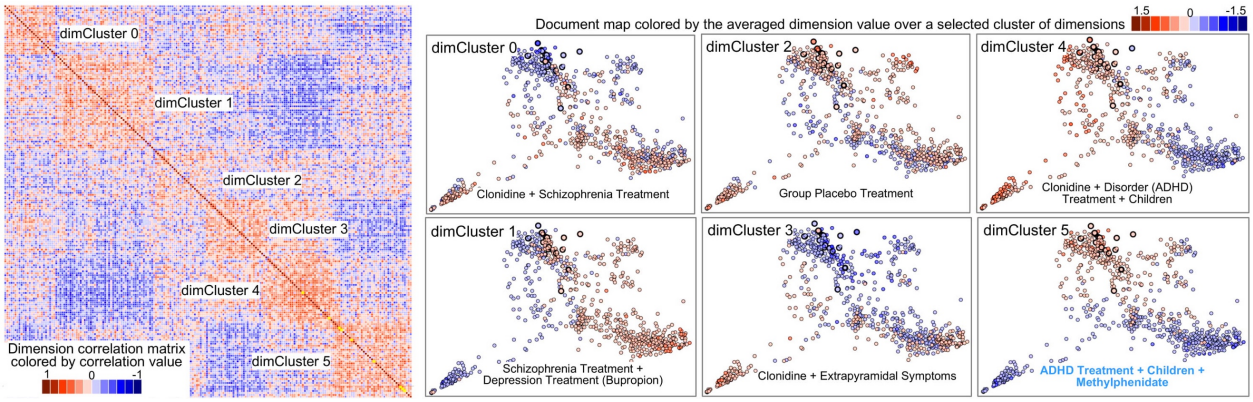


Fig. 7. Left: Dimension correlation matrix with clusters of correlated dimensions. Right: Adapted document map and semantics for each of the dimension clusters; the document map is colored by the averaged dimension value over a specific cluster of dimensions.

of variability can be used, e.g., ~SD and ~PCA. Users can also rank dimensions by the mean (average) value in Cx.

With a range of ranked dimensions in PCs, users can click any one dimension for validation and reasoning. As mentioned in Section 5.1.4, for a specific dimension d, the document map can adapt to reflect its value distribution, such that each document x is colored by the dimension value $v_{x,d}$ with a heatmap scheme. *Responsive or representative documents* (e.g., with a high positive value in the dimension) are also used for topic synthesis and to infer the dimension's semantics. As shown in Figure 5(b1)-(b4), D91, D166, D42, and D94 from the top list can differentiate C2 (*ADHD treatment*) from the others to a certain extent. D166, D42, and D94 are probed to encode semantics relevant to C2, in a general, specific, or mixed way. In contrast, in Figure 5(d1)-(d2), D128 and D55 from the bottom list have almost non-differentiating values across the map.

Users can feed multiple or all dimensions (V) from the displayed list to the feature selection box, which might already have a set of selected dimensions (U), with 3 options: union ($U \cup V$), intersection ($U \cap V$), or deduction ($U - V$). For example, users can select dimensions of high coherence in Cx, and then intersect with dimensions of high variability across all clusters. This could result in dimensions with both intra-cluster coherence and inter-cluster difference.

5.3 Document-Level Dimension Behavior

Similar to Section 5.2, we visualize dimension behaviors across a set of recognized documents. The documents (i.e., *reference documents*) can be benchmarks or new discoveries per domain and task interest. Compared to the semantics conveyed by a recognized cluster, reference documents could imply more specific semantics for IR. Besides reference documents, users can also specify *contrast documents*, which can be irrelevant or sampled ones. As suggested by domain experts, we pre-selected a contrast set with representative documents from all clusters across the corpus.

As shown in Figure 6, we use heatmaps to visualize dimension values across a set of documents, so that documents are arranged in rows, dimensions are arranged in columns, and dimension values are encoded by colors (same color legends with Figure 5). The design rationale lies in that heatmaps can provide side-by-side comparisons across documents, while PCs are better for scalability and accumulated patterns. We also had a prototype using another set of PCs, with axes for documents and polylines for dimensions. But feedback suggested that users might get confused by two different sets of PCs (Section 5.2.1).

Heatmap columns (dimensions) can be ranked by a dimension's document-level behaviors: 1) variability (SD), 2) coherence (~SD), and 3) mean (average) value across the selected documents; and 4) the dimension value in one specified document. The first three options are similar to Section 5.2.2; and the fourth option is to examine the alignment to the specified document. Figure 6 shows a better alignment across 5 benchmark relevant documents (with some noise), compared to 7 sampled contrast documents. Users can decide the range of dimensions to be displayed, and validate and select dimensions as in Section 5.2.2.

5.4 Visualization of Dimension Correlation

Sections 5.2 and 5.3 allow users to explore and identify dimensions in a guided manner. Here we visualize dimension correlation for users to explore the inherent patterns and expand identified dimensions with correlated ones.

As shown in Figure 7, with n (e.g., n=200) neural dimensions in the embedding space, we visualize an n*n correlation matrix, where each cell represents and is colored by the correlation between the row and column dimensions. To reveal clusters of correlated dimensions, we enable either hierarchical or *k-means clustering (default and k=6)*, applied to either dimension values in documents or *dimension correlation values (default)* [39]. The results are used to arrange dimensions from the same cluster into adjacent rows and columns, thus clusters of correlated dimensions are revealed as blocks of highly valued cells along the diagonal.

Users can search identified salient dimensions, and the corresponding cells (row + column matches) in the matrix will be highlighted along the diagonal, indicating the corresponding clusters. As shown in Figure 2(d), users can observe that salient dimensions with similar behaviors tend to distribute into the same cluster(s). Users can also feed a cluster of correlated dimensions to feature selection, and investigate the adapted document map, which is colored by the averaged dimension value over selected dimensions (Section 5.1.4). Topic synthesis will also be triggered to infer the semantics of a cluster of correlated dimensions. Figure 7 shows that different clusters can account for different patterns across the corpus, such as the value distribution, discriminative capability, and associated semantics.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2903946, IEEE Transactions on Visualization and Computer Graphics

6 USE CASES

6.1 ADHD Treatment

E1 was involved in an IR task to identify relevant clinical trials from the ADHD corpus to inform *ADHD treatment efficacy* (information need) [29]. This corpus (851 studies) is from a completed SR with 20 studies labeled as benchmark relevant ones. E1 was interested in how domain meaningful and task important semantics (biomedical and clinical concepts) were encoded, and how that contributed to IR (*P: Precision, R: Recall*). This case is demonstrated in Figure 2.

E1 was provided with a document map with clusters and topics indicating different types of mental illness and treatments, as shown in Figure 2(a). E1 confirmed the usefulness and meaningfulness of this overview, and rapidly found the green cluster (C2: *ADHD treatment*) to be the most relevant **[T1]**. E1 also noticed 19/20 benchmark relevant studies were in C2, indicating a baseline performance of R=95% (19/20) and P=5.40% (19/352), where 352 is the size of C2. While the embedding consisted of n=200 dimensions, E1 wanted to identify dimensions associated with *ADHD treatment*. For doing so, she had a top-down exploration guided by C2, invoked PC visualization, and ranked dimensions by their capability of differentiating C2 from the others, as shown in Figure 2(b). She then examined the top 30 dimensions and commented that the visual encodings (e.g., aggregated polylines) were effective to reveal dimension behaviors contributing to C2 **[T2]**. E1 applied all the 30 dimensions to feature selection by re-generating the document map and clusters in the new space, as shown in Figure 2(f1). Surprisingly, all 20 benchmark relevant studies were captured in the same cluster with a size of 342, which means that with only 30 salient dimensions directly, the recall was improved from 95% to 100%, even with a slightly improved precision from 5.40% to 5.85% (20/342) **[T6]**. For another top-down exploration, E1 recognized 5 relevant documents (references) in C2, invoked heatmap visualization, ranked dimensions by coherence across the references (Figure 2(c)), and selected the top 50 dimensions (U) **[T3]**. More meticulously, she deducted U by 100 dimensions (V) with high coherence across contrast documents. By applying the resulting 16 dimensions (U-V) to feature selection and using a distance threshold to the references, 189 responsive documents were captured, including all 20 benchmark documents, as shown in Figure 2(f2). E1 was impressed that with only 16 dimensions, the precision was improved from 5.40% to 10.58% (20/189) with a 100% recall **[T6]**. E1 also took some dimensions from T2 and T3 for closer examinations and gained interesting observations (Figure 2(e1)-(e4)): neural dimensions can encode semantics from general to specific, forming an evolution from *patient treatment*, *group placebo treatment*, *disorder treatment*, to *ADHD treatment on children with methylphenidate* **[T5]**. Finally, E1 explored dimension correlation and found that 9 salient dimensions (peak positive value) from T2 were distributed into 2 of the 6 dimension clusters; dimCluster#5 alone covered 7 of the 9 salient dimensions **[T4]**, as shown in Figure 2(d). E1 applied all dimensions from dimCluster#5 to feature selection. With a threshold for the averaged dimension value, a precision of 6.71% (20/298) and recall of 100% was achieved **[T6]**, as shown in Figure 2(f3). E1 appreciated that simply using a favored cluster of correlated dimensions, the baseline performance was improved.

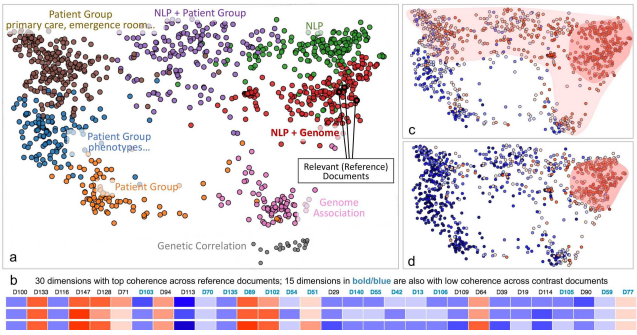


Fig. 8. Document map, identification of salient dimensions, and feature selection on Patient Phenotype Cohort Identification corpus.

Overall, E1 was impressed by the exploration process to gain insight and confidence in the neural document embedding space. She highly valued the improved IR performance achieved by selecting a small set of favorable dimensions (semantic features) in an advisable way.

6.2 Patient Phenotype Cohort Identification

E2 worked on a corpus consisting of 1044 biomedical publications regarding *patient phenotyping and cohort identification*. He aimed to explore encoded domain semantics and identify studies about NLP in his clinical research area. E2 appreciated the meaningfulness and interestingness of the document map (Figure 8(a)), and commented the clusters/topics well characterized his area, involving *patient groups and phenotypes*, *genetics*, *NLP in cohort identification*, etc. **[T1]**. E2 then investigated the red cluster with the semantics of *NLP + genome* and rapidly recognized 3 relevant documents. E2 decided to have a guided exploration to look for salient dimensions associated with the documents (references). He applied the top 30 dimensions with high coherence across the references to feature selection, but found they had limited discrimination over the corpus (Figure 8(c)). Thus, he deducted them by another set of dimensions with high coherence across contrast documents, and obtained 15 remaining dimensions as highlighted in Figure 8(b) **[T3].** As shown in Figure 8(d), by applying the 15 dimensions to feature selection, E2 received an adapted document map colored by document similarities to the references, identifying 81 documents (in dark red) highly similar to the references. E2 scanned the retrieved documents and was satisfied with their spectrum and quality, towards *integrating clinical genome research to promote domain NLP* **[T6]**. Overall, E2 appreciated the improved analytics that helped to understand a corpus in a neural embedding space, and was satisfied with the guided feature selection that helped to identify relevant studies of interest.

7 DISCUSSION

Features, documents, and topics. There are three granularities of data used in this study: features (dimensions), documents, and topics (clusters or groups of documents). Features are low-level building blocks of documents, and topics are high-level synthesis of documents sharing similar

properties. We use documents and (interpretable) topics to probe semantic meanings and analytic facets, and guide an exploration of (abstract) features or neural dimensions.

Semantics of neural document embedding. We confirmed that a neural document embedding space can consist of diverse information, noise, and redundancies. Some neural dimensions encode semantics of domain and task interest, ranging from general concepts (e.g., *disorder treatment*) to specific concepts (e.g., *ADHD treatment with methylphenidate on children*), or with a mixture of multiple concepts (e.g., *ADHD treatment + clonidine*). There are also dimensions with non-differentiating or mixed values across the corpus accounting for noise. Also, there are groups of correlated dimensions with similar behaviors. Users can benefit from such insights to exploit dimensions (or group of correlated dimensions) encoding favored semantics, instead of dimensions with unfavored semantics or noise.

Semantics and dimensions. Considering the distributed nature and multi-to-multi relations between semantics and dimensions, certain semantics can be encoded by multiple dimensions. Our current scope more focuses on *probing* semantics of IR interest and *exploiting* a set of multiple associated or salient dimensions, via ranking and flexible selection (unweighted combination). It is a future scope to have a linear (weighted) combination of dimensions that may best *express* or *characterize* certain semantics. Prospectively, users would adjust (low-level) weights or contributions of selected dimensions in an iterative and advisable manner; or input (high-level) task specifications, e.g., annotations or labels implying certain semantics, and trigger a training of the weights assigned to dimensions. On the other hand, a dimension can encode mixed semantics as afore-discussed. Therefore, instead of assigning exclusive meanings to individual dimensions, we use topic synthesis on representative documents to bridge encoded semantics that can be expressed in a noisy, mixed, or specific (explicit) way. In summary, while it is difficult to assign absolute roles regarding the complexity of natural languages and semantic representations, there is an information seeking need to approach *more relevant or favorable* semantics, and exploit dimensions with *more associated or explicit meanings*.

IR with human in-the-loop. Many IR applications rely on task specification to identify relevant documents, but prior supervision can be expensive and less generalizable. Visual analytics with feature selection contributes a new venue for that. In a guided, top-down, and interactive manner, users can identify and apply salient features (dimensions) based on task interest or their own knowledge, and retrieve relevant documents in an effective and efficient way. This aligns with the research trends to keep human in-the-loop, allowing knowledge injection, refinement, and steering.

Usability. Domain experts agreed that the system functions were well designed and implemented to support various types of explorations on demands. They also valued the usefulness of the provided visual representations and interactions. A video tutorial has been helpful to provide a step-by-step guide. As the system interface consists of multiple components to support numerous visual analytics tasks, the current interface could be improved for its learnability. The feedback we received has inspired us to stratify our interface for different types of users, for example, a more simplified and contextual layout for general users (IR users); and a more controllable layout for experienced users (IR designers). Our future work also includes continuing the iterative design and evaluation with other end-users to keep on the improvements.

Generalization and scalability. Built upon vector representations, our system can be generalizable to document embeddings produced by other models. With the guided exploration and ranking of neural dimensions, our system can be scalable to a higher-dimensional embedding space.

Analyzing a neural embedding model. Instead of PV's training dynamics, we focused on PV's final states for two main reasons: (1) neural document embeddings are considered the final states of the hidden layer and are widely used in applications, and (2) training dynamics is usually studied for individual documents or sentences. With an application-driven view, we are interested in established embeddings across a corpus. In the future, we plan to incorporate an NLP-driven view to analyze learning dynamics and the parameters (e.g., window size and embedding size).

8 CONCLUSION

We designed and developed a visual analytics system to explore neural document embedding, interpret the semantic properties, and exploit semantic features (neural dimensions) to promote IR applications. The system components include a configurable document map to provide throughout guidance and reasoning; views of global, cluster-level, and document-level dimension behavior to identify salient dimensions; and a view of dimension correlation to identify correlated dimensions. Together, they enable top-down exploration and bottom-up reasoning. We conducted use cases on real-world datasets with inspiring findings: neural embedding encodes diverse semantics and noise; a small subset of dimensions encoding semantics of domain and task interest can bring improved IR performance; a selected group of correlated dimensions can benefit IR directly, etc. We received positive feedback from domain experts verified the system's usefulness and effectiveness.

ACKNOWLEDGEMENT

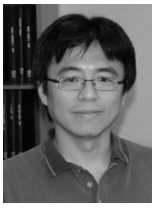
This work was supported in part by the Agency for Healthcare Research and Quality (AHRQ), R03HS025047-01. The authors would like to thank Albert Lai, Junpeng Wang, and anonymous reviewers for their generous help.

REFERENCES

[1] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.
[2] A. M. Dai, C. Olah, and Q. V. Le, "Document Embedding with Paragraph Vectors," *arXiv Prepr. arXiv*, vol. 1507, no. 07998, 2015.
[3] M. M. Lopez and J. Kalita, "Deep Learning applied to NLP," *arXiv Prepr. arXiv*, vol. 1703, no. 03091, 2017.
[4] B. Mitra and N. Craswell, "Neural Text Embeddings for Information Retrieval," in *ACM - WSDM '17*, 2017.
[5] F. Hill, K. Cho, and A. Korhonen, "Learning Distributed Representations of Sentences from Unlabelled Data," *arXiv Prepr. arXiv*, vol. 1602, no. 03483, 2016.

[6] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative Study of CNN and RNN for Natural Language Processing," *arXiv Prepr. arXiv*, vol. 1702, no. 01923, 2017.
[7] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *ECCV*, 2013.
[8] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv Prepr. arXiv*, vol. 1312, no. 6034, 2013.
[9] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards Better Analysis of Deep Convolutional Neural Networks," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 91–100, 2017.
[10] P. E. Rauber, S. G. Fadel, A. X. Falcao, and A. C. Telea, "Visualizing the Hidden Activity of Artificial Neural Networks," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 101–110, 2017.
[11] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and Understanding Recurrent Networks," *arXiv Prepr. arXiv*, vol. 1506, no. 02078, 2015.
[12] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and Understanding Neural Models in NLP," *arXiv Prepr. arXiv*, vol. 1506, no. 01066, 2015.
[13] V. Cirik, E. Hovy, and L.-P. Morency, "Visualizing and Understanding Curriculum Learning for Long Short-Term Memory Networks," *arXiv Prepr. arXiv*, vol. 1611, no. 06204, 2016.
[14] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush, "LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 667–676, 2018.
[15] R. Kiros *et al.*, "Skip-Thought Vectors," in *Advances in neural information processing systems*, 2015, pp. 3294–3302.
[16] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, "Embedding Projector: Interactive Visualization and Interpretation of Embeddings," Nov. 2016.
[17] H. Palangi *et al.*, "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 4, pp. 694–707, 2015.
[18] Z. Lin *et al.*, "A Structured Self-attentive Sentence Embedding," *arXiv Prepr. arXiv*, vol. 1703, no. 03130, Mar. 2017.
[19] S. Liu *et al.*, "Visual Exploration of Semantic Relationships in Neural Word Embeddings," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 553–562, 2018.
[20] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush, "SEQ2SEQ-VIS : A Visual Debugging Tool for Sequence-to-Sequence Models," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 353–363, 2019.
[21] B. C. Kwon *et al.*, "RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records," *IEEE Trans. Vis. Comput. Graph.*, May 2018.
[22] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. P. Chau, "ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 88–97, Jan. 2018.
[23] F. Heimerl, M. John, Qi Han, S. Koch, and T. Ertl, "DocuCompass: Effective exploration of document landscapes," in *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2016.
[24] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "FacetAtlas: Multifaceted Visualization for Rich Text Corpora," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1172–1181, 2010.
[25] X. Ji, R. Machiraju, A. Ritter, and P.-Y. Yen, "Examining the Distribution, Modularity, and Community Structure in Article Networks for Systematic Reviews.," *AMIA Symp.*, 2015.
[26] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist, "TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 151–160, 2017.
[27] X. Ji, R. Machiraju, A. Ritter, and P.-Y. Yen, "Visualizing Article Similarities via Sparsified Article Network and Map Projection for Systematic Reviews.," *Stud. Health Technol. Inform.*, 2017.
[28] T. Ruppert *et al.*, "Visual Interactive Creation and Validation of Text Clustering Workflows to Explore Document Collections," *Electron. Imaging*, vol. 2017, no. 1, pp. 46–57, 2017.
[29] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, "Reducing workload in systematic review preparation using automated citation classification," *J. Am. Med. Inform. Assoc.*, 2006. https://dmice.ohsu.edu/cohenaa/systematic-drug-class-review-data.html
[30] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," *arXiv Prepr. arXiv*, vol. 1805, no. 01070, May 2018.
[31] T. Schreck, J. Bernard, T. Tekusova, and J. Kohlhammer, "Visual cluster analysis of trajectory data with interactive Kohonen Maps," in *2008 IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 3–10.
[32] D. Sacha *et al.*, "Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 241–250, Jan. 2017.
[33] H. Kim, J. Choo, H. Park, and A. Endert, "InterAxis: Steering Scatterplot Axes via Observation-Level Interaction," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 131–140, Jan. 2016.
[34] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, "Dis-function: Learning distance functions interactively," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012.
[35] M. Yoshioka, M. Itoh, and M. Sebag, "Interactive Metric Learning-based Visual Data Exploration: Application to the Visualization of a Scientific Social Network," in *ISIP*, 2015.
[36] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Min. Appl. Theory*, vol. 1, no. 20, 2010.
[37] Y. Zhang, X. Ji, M. Ibaraki, and F. W. Schwartz, "Mining Information from Collections of Papers: Illustrative Analysis of Groundwater and Disease," *Groundwater*, 2018.
[38] M. Tory, C. Swindells, and R. Dreezer, "Comparing Dot and Landscape Spatializations for Visual Memory Differences," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1033–1040, 2009.
[39] J. Wang, X. Liu, and H.-W. Shen, "High-dimensional data analysis with subspace comparison using matrix visualization," *Inf. Vis.*, 2017.

Xiaonan Ji received the BE degree in Computer Software Engineering from Beihang University, and the MS and PhD (2018) degrees in Computer Science and Engineering from The Ohio State University. She is a Postdoc with the Institute for Informatics, Washington University School of Medicine. Her research interests include visual text analysis, NLP, information retrieval, and clinical informatics.



Han-Wei Shen received the BS degree from the Department of Computer Science and Information Engineering, National Taiwan University, the MS degree in Computer Science from the State University of New York at Stony Brook, and the PhD degree in Computer Science from the University of Utah. He is a full professor with The Ohio State University. His primary research interests include scientific visualization and computer graphics.

Alan Ritter received the BS and MS degrees in Computer Science from Western Washington University, and the PhD degree in Computer Science and Engineering from University of Washington. He is an assistant professor with The Ohio State University. His primary research interests include natural language processing, information extraction, machine learning, and social media analysis.

Raghu Machiraju received his PhD degree in Computer and Information Science from The Ohio State University (OSU). He is a full professor at OSU in Biomedical Informatics, Computer Science and Engineering, and Pathology. He is the Executive Director of the Translational Data Analytics Institute. His primary research interests include visual analytics, modeling, and machine learning as applied to biological and life sciences.

Po-Yin Yen received the BS degree in Nursing from National Cheng Kung University, the MS degree in Medical Informatics from Oregon Health & Science University, and the PhD degree in Nursing from Columbia University. She is an assistant professor with the Institute for Informatics, Washington University School of Medicine, and Goldfarb School of Nursing, Barnes-Jewish College, BJC HealthCare. Her research focuses on human-computer interaction, workflow analysis, and data visualization to support clinical practice.