Visual Exploration of Semantic Relationships in Neural Word Embeddings

Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar,
Bei Wang, Yarden Livnat and Valerio Pascucci

Abstract— Constructing distributed representations for words through neural language models and using the resulting vector spaces for analysis has become a crucial component of natural language processing (NLP). However, despite their widespread application, little is known about the structure and properties of these spaces. To gain insights into the relationship between words, the NLP community has begun to adapt high-dimensional visualization techniques. In particular, researchers commonly use t-distributed stochastic neighbor embeddings (t-SNE) and principal component analysis (PCA) to create two-dimensional embeddings for assessing the overall structure and exploring linear relationships (e.g., word analogies), respectively. Unfortunately, these techniques often produce mediocre or even misleading results and cannot address domain-specific visualization challenges that are crucial for understanding semantic relationships in word embeddings. Here, we introduce new embedding techniques for visualizing semantic and syntactic analogies, and the corresponding tests to determine whether the resulting views capture salient structures. Additionally, we introduce two novel views for a comprehensive study of analogy relationships. Finally, we augment t-SNE embeddings to convey uncertainty information in order to allow a reliable interpretation. Combined, the different views address a number of domain-specific tasks difficult to solve with existing tools.

1 Introduction

Natural language processing (NLP) is one of the key components in today's digital world responsible for everything from web search to document classification and from machine translation to speech recognition. A crucial breakthrough that led to the recent surge of AI research in NLP is the concept of neural word embeddings, such as word2vec [27] or Glove [33]. These systems utilize a large corpus of training articles to determine the co-occurrence statistics between pairs of words within a given context, and employ a neural network to infer a vector space for embedding words. Interestingly, the position and difference vectors between words appear to encode semantic relationships (see Fig. 2). One of the most striking examples is analogy pairs such as (*king*, *queen*) and (*man*, *woman*). In the word embedding space, one finds that (*woman* + *king* - *man*) $\approx$ *queen* [29]. Broadly speaking, encoding words or even sentences into intermediate vector representations provides the foundation for a number of NLP applications, such as sentiment analysis [25,40] or document ranking [16]. However, despite its central importance and wide-scale adoption, the word embedding space remains a rather abstract and unintuitive concept to most NLP researchers.

To encode a large number of semantic relationships between words from a large corpus of text, the embedding dimension is chosen to be reasonably high ($\sim$300). Reasoning in such spaces is difficult and thus some NLP researchers have turned to visualization for more intuitive interpretations of the word embedding space. In particular, nonlinear dimension reduction strategies, most notably t-distributed stochastic neighbor embeddings (t-SNE) [42], are used to provide a high-level overview of the embedding space. Although such an embedding can reveal some interesting separation between word groups, i.e., *countries*, *nouns*, *verbs*, etc., they inherently distort the linear (semantic) relationships most interesting to researchers. Consequently, to preserve such relationships, linear projections are preferred. The most common approach is to use principal component analysis (PCA) restricted to carefully chosen subsets of words, i.e., *countries* and *capitals*, *nouns* and their *plurals*, etc. Unfortunately, both the linear (PCA) and nonlinear (t-SNE) approaches, which are now the de facto standard in NLP research, are fairly limited and often misleading. For example, t-SNE embeddings are often used to validate (or discredit) various intuitions on the nature of the embedding space without any consideration for the inherent distortions in the projection itself. Given the complex nature of the high-dimensional space, any two-dimensional embedding will exhibit significant distortions and thus any given feature may in fact be an artifact. Similarly, the PCA embeddings rely on the fact that the semantic direction of interest, i.e., the vector (*man* - *king*), has more variation than other directions. As demonstrated in this paper, such an assumption is sometimes true, in which case PCA embeddings work reasonably well. However, in other cases the variation within one word group, i.e., *countries*, can be greater than the distance to a related group, i.e., *currencies* and the PCA embedding fails to provide the expected alignment. In addition, for a given analogy type, binary labels for words are known. However, such important information is not utilized in the PCA. In general, we find that the embeddings used in NLP research are not necessarily ideal. Furthermore, in a number cases, embeddings are provided as is, with little information on how they were created or how reliable they might be. This lack of information invariably leads to misuse or misleading interpretations of the visualization results.

As part of a long-standing collaboration with domain scientists, we present a system aimed at addressing some of these problems. The goal is twofold: First, to develop new tools specifically designed to answer such questions as how well a given semantic relationship is approximated by a single direction or how different semantic concepts are related (tasks the proposed tool aims to address are characterized in Table 1); and second, to provide users with more information about how to interpret the visualization results. In particular, we enhance the global view (computed by t-SNE) by incorporating per-word distortion metrics as well as an interactive display of neighboring words in the high-dimensional space. This augmentation provides an intuitive illustration of which apparent features in the data are trustworthy and a quick way to explore the embedding in detail. Furthermore, we introduce new approaches to compute linear embeddings of semantic relationships (by utilizing the binary label information in an analogy type) that simultaneously maximize the separation of the two concepts, i.e., *male* vs. *female*, and minimize the differences between semantic directions,

- *Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan is with Lawrence Livermore National Laboratory. E-mail:{liu42,bremer5,jayaramanthi1}@llnl.gov.*
- *Bei Wang, Yarden Livnat and Valerio Pascucci is with SCI Institute, University of Utah. E-mail:{beiwang,yarden,pascucci}@sci.utah.edu.*
- *Vivek Srikumar is with School of Computing, University of Utah. E-mail: svivek@cs.utah.edu*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

i.e., the vectors (*man* - *king*) vs. (*woman* - *queen*). Finally, to verify that the resulting projections indeed capture true semantics rather that an accidental alignment, we extend the notion of hypothesis testing to the embedding and provide users with the equivalent of a p-value for a given result. Our contributions in detail are:

- A characterization of domain-specific tasks for exploring semantic relationships in neural word embeddings (see Table 1);
- An interactive word embedding visualization tool specifically designed to support NLP research;
- A generalization of hypothesis testing for embeddings to evaluate the saliency of the apparent structure; and
- A case study demonstrating new insights into a well-known semantic analysis dataset.

2 Related Work

Visualization has been used to tackle several challenges in text analysis and NLP, such as topic modeling and sentiment analysis. In the existing literature, several visualization systems, including the *Termite* [8], the *Hiérarchie* [38], and the concurrent words and topics visualization [37], have attempted to include humans in the analysis loop for identifying topics through visual encodings. Sentiment analysis applications [10, 23, 24, 45] find and track the sentiment toward topics or summarize the evolution of topics in social media feeds and other mass text sources. These techniques are useful, but they focus on more abstract concepts and are not well suited to understand low-level details present in word embeddings such as analogy relationships. Furthermore, understanding neural models as well as their training process has also attracted interests. Interactive tools, such as LAMVI [35], are designed to help domain experts understand the effects of training parameters and debug the training process. The recent work of Li et al. [19] focuses on visualizing the compositionality (build sentence meaning from the meanings of words and phrases) of vector-based models. Gladkova et al. [12] questioned the existing intrinsic (semantics) evaluation approach for word embeddings that relying on abstract ratings and argued the importance of exploratory evaluation that characterizes embeddings' strengths and weaknesses. Compared to techniques often seen in the visualization community (where novel visual encoding plays a key role), the focus of these works [12, 19] is on carefully designed experiments that stem from an in-depth understanding of the model. The visualization (e.g., heatmap) mostly plays a supplementary role in aiding the interpretation of the experiment results.

There are generic dimension reduction tools that can be applied to word embeddings. For example, the *Embedding Projector* [36] is a new embedding visualization tool released by Google as part of the TensorFlow framework [2]. In addition, a number of openly available toolkits such as *scikit-learn* [32] implement several dimension reduction approaches. Currently, the t-SNE embedding [42] is the most commonly adopted approach for visualizing word embeddings. Compared to other common nonlinear dimension reduction techniques, t-SNE is optimized for 2D visualization and is more likely to reveal inherent clusters in the data. Consequently, it is often used to provide a quick overview of the overall structure and to highlight separation between word categories. However, since t-SNE creates nonlinear embeddings, the linear relationships most interesting to researchers are invariably lost. As a result, in practice, PCA of known subsets of words is used to visualize linear relationships. Since the vector values in the embedding space have no explicit meaning, many popular high-dimensional visualization techniques, such as parallel coordinates [14] and scatter plots matrices [7], are less appropriate. A related challenge regarding any 2D projection of the word embedding space is the error (uncertainty) invariably introduced during the dimension reduction process. However, these challenges are rarely considered or visualized explicitly in the NLP community, which presents a risk for gross misinterpretation of the data. Here, we adopt the embedding quality measures [6, 13, 17, 41, 43] previously explored in the visualization community [21, 26, 30] to aid in the interpretation of uncertainty in the t-SNE embeddings. For more information on uncertainty visualization techniques, we refer the reader to the surveys in [5, 34].

Nevertheless, none of the techniques discussed above address the specific needs of NLP researchers nor do they provide capabilities beyond what is currently state of the art in the NLP community. After an extensive literature search, to the best of our knowledge, a dedicated tool that helps NLP researchers understand and explore high-dimensional word embeddings does not exist.

3 Background: Neural Word Embeddings

From web search to voice recognition, the advances in NLP have shaped how we interact with the digital world. At the core of several modern NLP systems is the concept of neural word embeddings, such as word2vec [27] and Glove [33], which provide a powerful, distributed representation for words [16, 40]. In particular, the embeddings support any number of higher level analysis tasks, such as sentiment analysis [25, 40], machine translation [46], and document modeling [16].

The general idea behind word embeddings can be described as follows (see Fig. 1): Let us assume we have a vocabulary of n words $\{w_1, \ldots, w_n\}$ extracted from a large text corpus (e.g., Wikipedia). We first create the co-occurrence statistics matrix $\mathbf{M}$ (e.g., pairwise mutual information) in which each entry $M(i, j)$ encodes how strongly w_i and w_j are related. Loosely speaking, one might interpret $M(i, j)$ to encode the probability for words w_i and w_j to appear together within the same context. Subsequently, we can employ a variety of optimization strategies, such as Glove or skip-gram with negative sampling, to obtain a metric space for words that preserve the relationships encoded in $\mathbf{M}$. Interestingly, it was showed in [18] that the word embedding optimization is equivalent to performing matrix decomposition (symmetric SVD) on the matrix $\mathbf{M}$ directly.

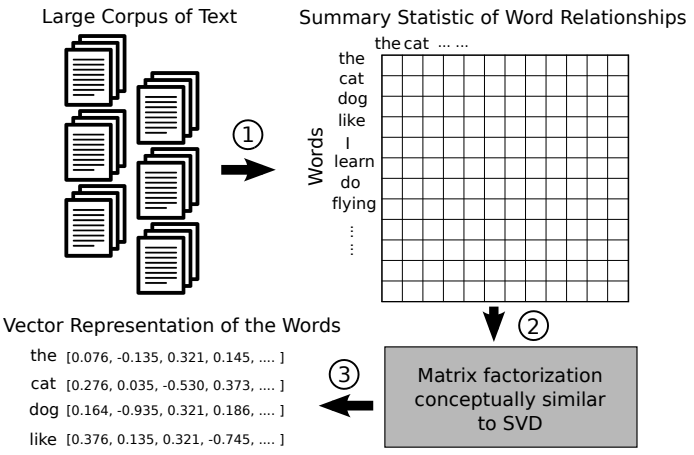


Fig. 1. An illustration of the word embedding process: The input of the algorithm is a large corpus of text, which is summarized in a $n \times n$ matrix M that encodes the relationships between n unique words. Typically, $M(i, j)$ records statistical relationships, such as the probability of joint occurrence between $word_i$ and $word_j$. Subsequently, M is factorized and the coordinates in the $d \ll n$ most significant components define the vector representation of words.

In order to obtain a quantitative understanding of the word embeddings, it is common to analyze the vector difference between word vectors. More specifically, word positions and difference vectors encode crucial semantic and syntactic information. One of the surprising findings is that in the learned vector space, words support simple, algebraic manipulations. A prototypical example is the study of analogy pairs – *king*:*queen*, *man*:*woman* – where *king* - *man* + *woman* is approximately equal to *queen* (see Fig. 2). The analogy relationships have proven so useful that they are now routinely used to evaluate how well an embedding is capturing the semantic and syntactic characteristics [29]. Nevertheless, due to the high-dimensional nature of the vector space, researchers still have only a limited understanding of the true relationships between words.

Table 1. A list of prominent NLP tasks pertinent to word embeddings, gathered from NLP experts. We identify which of these tasks can be performed using existing solutions (adopted in NLP community) and highlight the gaps bridged by the proposed approach .

	Tasks	Existing	Proposed
1	What is the overall distribution of words or clusters?	✓ [42]	✓
2	Can we evaluate the quality of neighborhood preservation in a 2D embedding?	✗	✓
3	How can we view high-dimensional neighborhood information in a 2D embedding?	✓ [36]	✓
4	Can we find the most dominant linear structure for a given analogy relationship?	✓ [15]	✓
5	Can we find a linear projection that highlights analogy relationships?	✗	✓
6	Are certain analogy relationships observable only in subspace?	✗	✓
7	Can we identify the dimensions directly corresponding to semantic concepts (e.g., masculine → feminine)?	✗	✓
8	Are there subtrends within an analogy relationship?	✗	✓
9	How can we visually compare different word embeddings?	✗	✓

Finally, we define a few NLP-related terms used in the paper for clarity. An *analogy pair* is used to indicate a pair of words exhibiting a specific analogy relationship (e.g., *man*:*woman*). An *analogy group* is a set of analogy pairs sharing the same analogy concept and an *analogy direction* is used to denote the difference vector, *man* - *woman*.

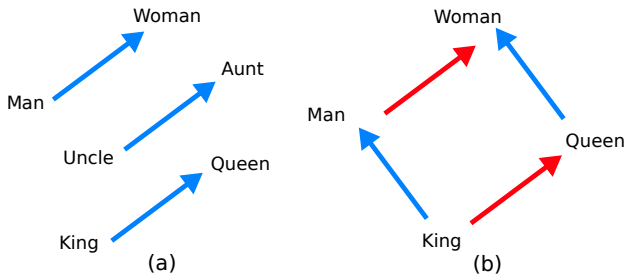


Fig. 2. In the word embedding space, the analogy pairs exhibit interesting algebraic relationships.

4 Design Process

The project started from an incidental demonstration of a high-dimensional visualization tool to an NLP researcher – one of the coauthors. The NLP expert remarked that such a system had the potential to be very helpful in his line of research and began introducing us to the challenges of word embeddings. We subsequently recruited additional NLP experts collaborating with us on other projects and jointly started to discuss the specific visualization needs of the NLP community. Through an iterative process of design, development, and evaluation, the team specialized the tool for NLP problems. Some of the important milestones in this process were the realization that most questions of interest involve only a small subset of words, i.e., a single analogy category, and that the subjective notion of which embedding is the most informative is not necessarily connected to any of the commonly used metrics of distortion. Furthermore, visualization practices in some of the well-known NLP publications seemed problematic from our perspective as visualization researchers. In particular, projection results are typically presented as is without assessing the impact of potential errors/uncertainty in the visual representation. In general, there is a lack of dedicated tools for answering the specific questions of greatest interest to the domain expert. We believe the NLP community could benefit from the introduction of better visualization tools tailored specifically for this domain.

Throughout the design process, we have assembled a list of specific visualization tasks (provided in Table 1), along with an initial assessment of whether these tasks can be addressed using tools that have already been adopted in the NLP community. As mentioned above, creating an embedding for all words (T1) addresses only a small aspect of the problem as it is clear that such a projection will inherently create severe clutter and large distortions. However, these embeddings are used to provide an overall context and our collaborators suggested that augmenting such embeddings with distortion measurements (T2) and the ability to explore neighborhoods of words (T3) would be useful. Most of the remaining tasks focus on using subsets of words. In particular, T4-T8 are all designed to explore analogy relationships as they are critical to understanding the semantics in word embeddings. The final task (T9) constitutes a generic goal to understand differences between the word embedding spaces produced by various methods, such as word2vec [28] and Glove [33].

5 Constructing Analogy Projections

The study of analogy pairs forms one of the basic building blocks in understanding word embeddings. In particular, researchers are interested in the separation between the concepts, i.e., splitting male vs. female terms, as well as in the analogy *directions*, i.e., the vector *king* - *queen*. Since the latter implies a linear relationship, nonlinear projections are not appropriate in this context. Instead, the NLP community is utilizing PCA in an attempt to highlight the prominent linear structure. However, PCA treats all words as individual points – not as analogy pairs – and simply captures the direction of the largest variation among all words. In many cases, the intergroup variance can be larger than the variance between the two concepts, resulting in an embedding that does not preserve the analogy relationship. For example, Fig. 4(a) shows the PCA for words with singular vs. plural analogy, which highlights word categories rather than the desired analogy direction. In general, a popular hypothesis among NLP researchers is that analogy pairs are linearly related only within certain (linear) subspaces. If this hypothesis holds, then unless the corresponding subspace represents the directions of dominant variation, PCA cannot provide a useful visualization. Furthermore, PCA, like most other common embedding techniques, ignores the separation between the two labeled concepts, which can lead to a poor analogy separation and unintuitive projections.

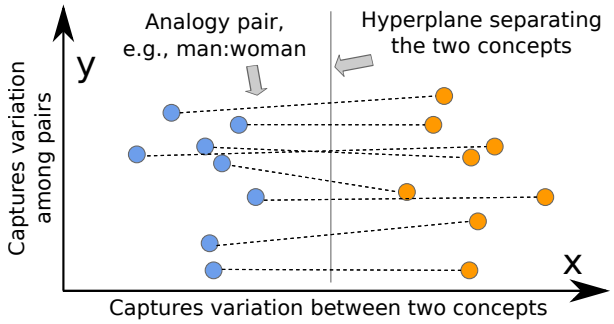


Fig. 3. An illustration of the intuition behind the projection-finding schemes for an analogy relationship. In a 2D projection, the x-axis captures variation between the two concepts in an analogy relationship. The y-axis captures the variation among pairs.

From discussion with our collaborators, two key objectives for an informative embedding emerged (illustrated in Fig. 3): First, the two concepts (male vs. female) should be well separated along one axis of the projection (here we choose the x-axis); and second, the different pairs, i.e., *king*:*queen*, *man*:*woman*, should preserve their relative distances in the orthogonal direction (in the y-axis of the projection). The first objective directly corresponds to a typical loss function in supervised classification, e.g., linear support vector machines [39] (SVMs),

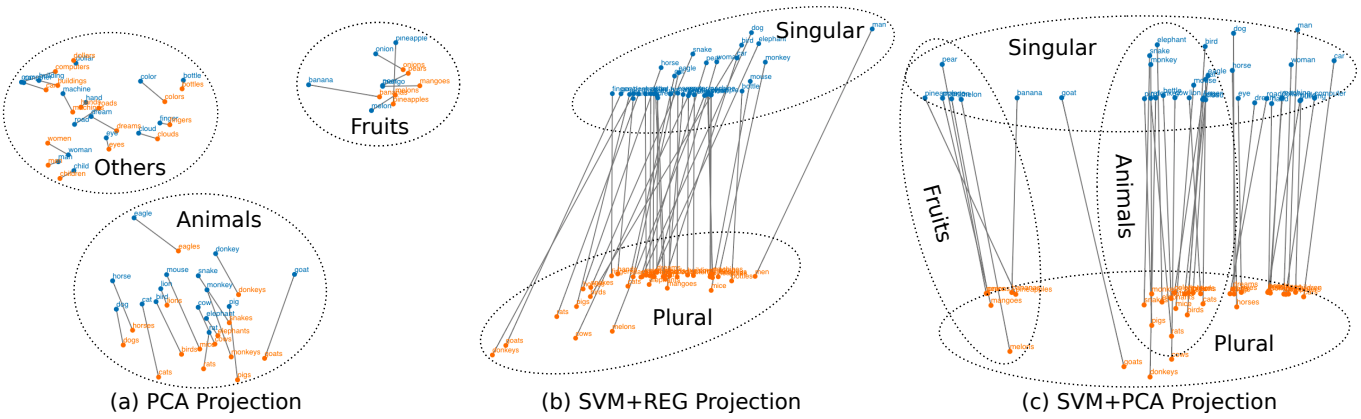


Fig. 4. The domain-specific projection-finding scheme. PCA (a) captures the largest variance in the data, which corresponds to the difference in the meaning of the nouns, whereas the proposed projection schemes (b) (c) capture the singular:plural analogy relationship. In this experiment, the official pretrained Glove vectors are used.

which attempts to find the hyperplane that best separates two classes of data. Consequently, we use the normal to the estimated hyperplane as the x-axis of our 2D embedding (see Fig. 3). We discuss how to generate the y-axis, which addresses the second objective, in the next paragraph. Since an analogy relationship has only a clearly defined binary label, linear discriminant analysis (LDA), despite being the default option for supervised 2D projection, does not apply directly. (For k-class dataset, LDA can produce only a k-1 dimensional embedding.) In addition, LDA assumes each class follows a normal distribution, whereas linear SVM does not make any assumption about the class distribution.

Here, we introduce two methods to optimize the y-coordinate of words (see Fig. 3). The first approach is designed to best align the analogy pairs, whereas the second focuses on preserving interclass distances. The former optimizes the y-axis direction to minimize the pairwise angles (in the 2D plane) between pairs. Given a set of analogy pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{T}$ ($\mathbf{x}_i$ is a word vector), the x-axis of the projection already captures the variation between the two concepts in the analogy relationship. Let $\mathbf{w}$ be the y-axis basis vector. To decrease the pairwise angles, we optimize $\mathbf{w}$ to make each pair as horizontal as possible, i.e., $(|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{y}_i| = 0)$. However, without additional regularization, the resulting embedding often collapses all words in each concept to a small area, which results in parallel lines indicating the analogy relationship, but does not produce an informative visualization. We resolved this issue by adopting an additional term in the objective function to require the projection to preserve the distances between words in the same concept, $\mathbf{x}_i - \mathbf{x}_j$. We pose this as a regression problem to predict the distances using the difference vectors. More specifically, we adopt the ℓ_2 regularized, ridge regression formulation: $\min_{\mathbf{w}} \sum_{p=1}^{P} \|d_p - \mathbf{w}^T\mathbf{v}_p\|_2^2 + \lambda\|\mathbf{w}\|_2$, where P denotes the total of pairs used for fitting the regressor, $d_p = 0$ if the difference vector $\mathbf{v}_p$ is constructed using words from an analogy pair and d_p is the actual Euclidean distance if $\mathbf{v}_p$ is constructed using two words from the same concept. Finally, the y-axis $\mathbf{w}$ is orthogonalized to the x-axis computed from linear SVM. In the rest of this paper, we will refer to this technique as SVM+REG. The results obtained using this approach for the singular vs. plural example are shown in Fig. 4(b). Compared to the PCA, the concepts are well separated (the result of the SVM) and the lines are roughly parallel.

The second approach, which is designed to better preserve intra-class distances, replaces the optimization of the first approach with a 1D PCA of word vectors from one concept. Since it better preserves distances within the same concept, this approach (referred to as SVM+PCA) often produces more intuitive arrangements and subjectively appears to better preserve interesting relationships. For example, Fig. 4(c) shows the results of the singular vs. plural embedding. Analogies are well separated and roughly parallel but again form multiple subgroups, which is meaningful with the left cluster representing fruits, the middle animals and the right cluster other words. The SVM+PCA approach, unlike the SVM+REG approach, does not explicitly optimize for parallel patterns among analogy pairs, which, in some cases, results in less consistency in creating a 2D visualization that maximally reveals the analogy relationships.

The interpretation of the projection is twofold. First, the proposed projection approaches made the assumption that the analogy relationship exists; therefore, it is crucial to verify whether the projection indeed captures the salient structure of the data instead of noise. In order to address such a challenge, we extend the concept of hypothesis testing to linear projection, where we test against the likelihood of finding a similar pattern in randomized data (where we are sure the analogy relationship does not exist). This test and the corresponding visual representation are an integral part of using the proposed projection method, as discussed in detail in Section 6. The second aspect of interpreting the plots is the kind of patterns we should look for in these projections. Since the goal of these projections is to highlight analogy relationships, the parallelism among lines connecting two words in an analogy is a good indicator of how strong the analogy relationship is, i.e., see Fig. 6(c)(d), where the same projection method is applied to two types of analogies. From the projections, we can see that (d) has a stronger analogy relationship than (c).

Finally, through extensive experiments and discussions with domain experts, it has become clear that a single projection cannot provide the user with all the important information to address all the tasks listed in Table 1. Therefore, instead of relying only on projections, we introduce two additional views (see Section 7.2) to provide a more comprehensive picture of the analogy relationships and to help address all tasks discussed in Section 8. In addition, the proposed system allows animation transitions (each frame is an in-between linear projection) among these different projections (PCA, SVM+PCA, SVM+REG), which provide additional structural insight into the word embedding space via exploratory analysis.

6 A Hypothesis Test for Projection Salience

The analogy projection of the previous section aims to find the linear subspace that optimally aligns the analogy relationship. However, given that we typically project only around 30 pairs (60 words) from 300D to 2D, it is possible that the large number of degrees of freedom produces artificially well-aligned vectors even for unrelated word pairs. Therefore, it is crucial to evaluate whether a well-aligned pair projection captures a truthful/salience analogy relationship or should be considered a false positive. In other words, we need to understand how reliable analogy projections represent the high-dimensional structure.

To guard against false positives, we adapt a standard hypothesis testing [3]. In particular, we test how likely it is that a certain projection result comes from random data. More specifically, our null hypothesis is that the current set of analogies has no correlation, i.e., the words are

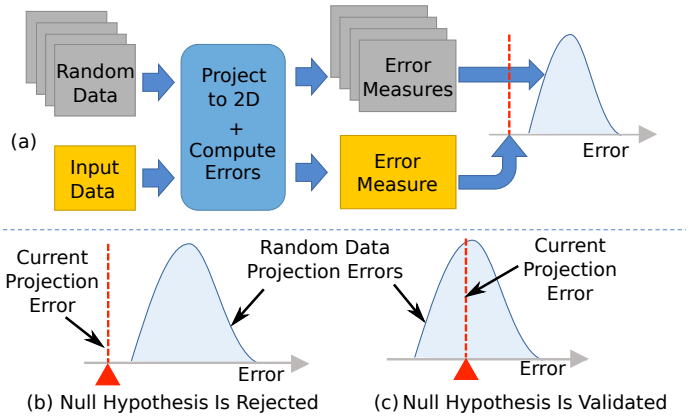


Fig. 5. Hypothesis testing for salient structure in linear projections. (a) During preprocessing (gray at the top), we project random pairs, compute their errors and assemble test statistics for the null hypothesis. Given an actual analogy pair as input, we display the error with respect to the test statistics. Projections that show salient structures are expected to have much lower errors than the test statistics (b). An error similar to or even larger than the test statistics (c) indicates that the projection is likely to contain artificial alignments.

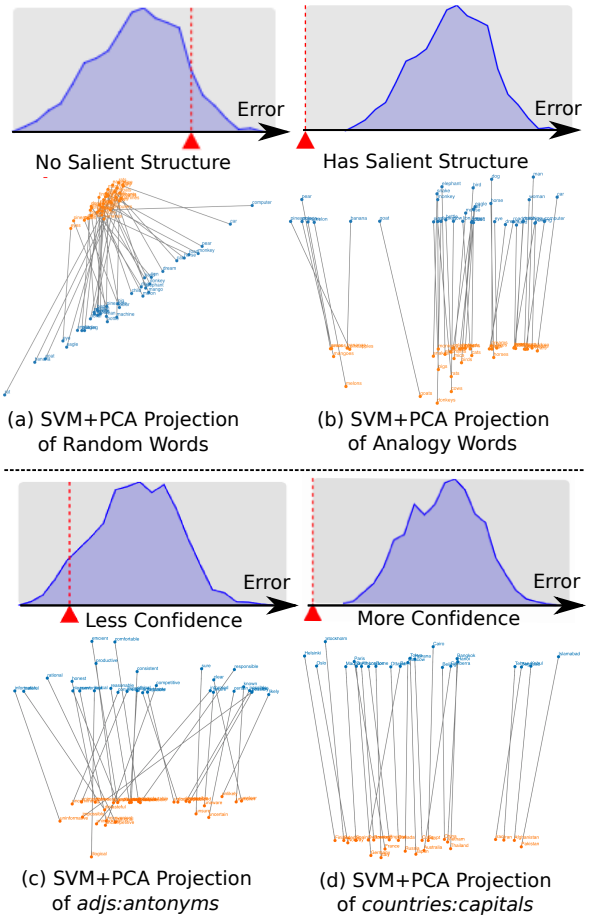


Fig. 6. Comparison of the hypothesis testing results between random data (a) and regular input (b), as well as between a strong analogy relationship (d) and a weak analogy relationship (c), using the same projection technique (SVM+PCA). The *analogy-direction error* plots tell us how likely it is the analogy relationship exists.

selected randomly from the word embedding. As a preprocessing step, we assemble a large number of random pairs of words, treat each of them as an analogy pair and embed them in 2D using analogy-projection techniques. For each 2D projection, we compute its *analogy-direction error*, defined as the sum of pairwise angles, an indicator of how well aligned all directions are in 2D. We then use these errors to estimate the distribution (the blue distribution in Fig. 5(a)) of the null hypothesis, which encodes how likely it is to see a given *analogy-direction error* from random data. Note that this distribution depends on the projection technique used, the dimension of the vector space and the number or pairs that are projected. In practice, we pre-compute distributions for a range of sizes.

Given a projection for word pairs, we compare the resulting *analogy-direction error* to the distribution of the error under the null hypothesis obtained with the same projection method, see Fig. 6. Such a plot provides a visual indication of the statistical significance (p-value [3]) for a given projection result, i.e., the likelihood for the strong correlations to be artifacts. As shown in Fig. 6(a), when the SVM+PCA projection optimized for the singular-plural analogy pairs is used with random words, the resulting *analogy-direction error* (red dotted line) overlaps with the null hypothesis distribution, thereby indicating the correlation pattern observed in the projection is artificial. On the other hand, for the actual analogy words (Fig. 6(b)), there is no overlap between the distribution and the *analogy-direction error*, which confirms the presence of salient structure. The test can also distinguish differences in confidence regarding the analogy relationship captured in the projection. As shown in Fig. 6(c)(d), based on the hypothesis-testing plots, we can see *countries:capitals* has a stronger analogy relationship compared to *adjs:antonyms-adjs*, which can also be observed from the projection.

Note that our null hypothesis is somewhat optimistic since random word pairs represent the worst-case behavior. As such, many of the projections of reasonably aligned pairs are significantly below the lower end of the null hypothesis distribution. Nevertheless, the relative distance still conveys the saliency of a given result in an intuitive manner, and we have encountered several cases in which optimized projections are well within the range of the null hypothesis. If needed, one could easily create additional hypothesis tests by, for example, selecting pairs between two randomly selected word categories.

7 Word Embedding Visual Explorer

Word Embedding Visual Explorer (see Fig. 7) is a web-based visualization tool for exploring the word embedding space. In this section, we describe the design choices, functionality, and the implementation of the system. To start the exploration, users can either select from a number of widely used pretrained word embeddings (e.g., Glove [33], Word2Vec [27]) or upload their own. Typically, users then load word groups of interest, for example, different analogy pairs or individual word categories. However, the exploration is not limited to the local scope among the words of interest. For certain tasks, the region of interest includes the entire word embedding space, in which case large numbers of word vectors will be fetched in the background and seamlessly passed to the analysis pipeline. The system consists of two major visualization components, the *Global t-SNE* view and the *Analogy Relationships* view. Both views are meant to replace the standard t-SNE and PCA with their enhanced counterparts.

7.1 Global t-SNE View

The global t-SNE view is designed to provide an overview of the arrangement of a large number of words. With respect to Table 1, the view aims to address tasks T1-T3 in the list. What differentiates this visualization from the standard t-SNE embedding is the ability to use visual encodings to aid the understanding of uncertainties and distortions in the 2D embeddings. According to a domain expert, users of t-SNE in NLP will often interpret inconsistency in the embedding, for example, a word far away from its expected neighborhood, as potential noise in the embedding, before considering the possibility of inaccuracies in the visualization. However, in our experience misplaced words are more likely the result of distortion in the dimension reduction. Since the intrinsic dimension of the words is expected to be $\gg 2$,

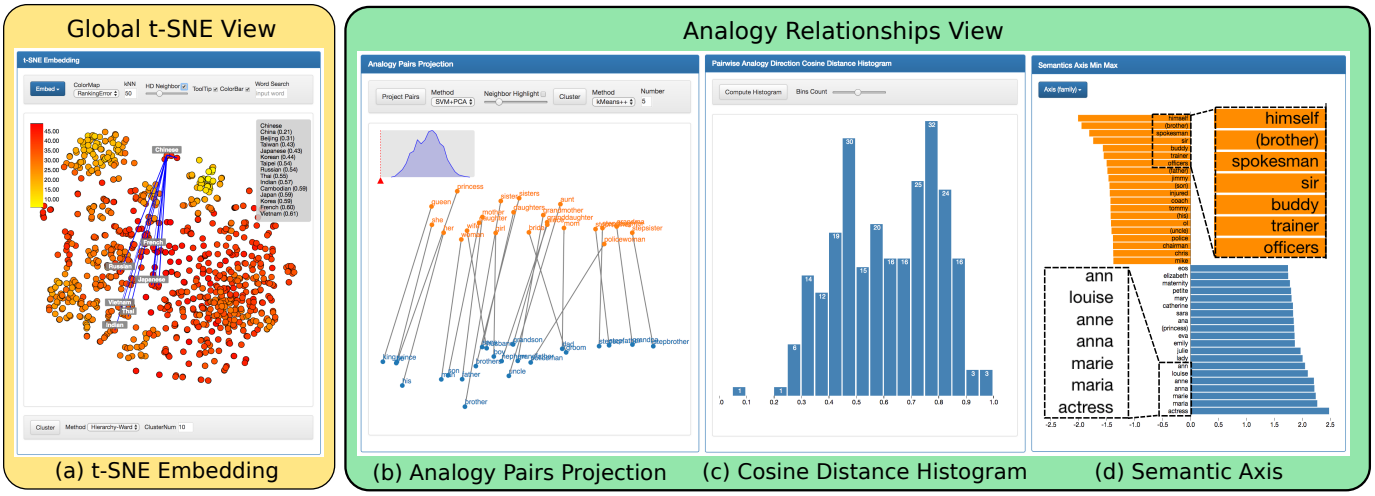


Fig. 7. Word Embedding Visual Explorer. (a) t-SNE embedding panel shows the overall structure of the words of interest. (b) Analogy projection panel enables the exploration among linear projections that highlight an analogy relationship. (c) Pairwise cosine distance histogram panel captures the overall parallelism of the analogy vector orientations. (d) Semantic axis panel shows the words at the extreme ends of an axis that capture a distinct concept.

such artifacts are unavoidable and should be taken into account before considering more complex explanations.

We adopt the concept of distortion error, in particular, per-point distortions [21, 30], to help illustrate the various levels of accuracy and reliability within the same embedding. To provide the per-point estimation, we compare how much a given point's neighborhood relationship changes in the 2D projection with respect to the original high-dimensional space. In this work, a ranking-based approach is adopted; the implementation details can be found in [17]. As illustrated in Fig. 7(a), points are colored based on how well their high-dimensional neighbors are preserved in the 2D space, allowing users to determine how much one should trust different regions of the same embedding. However, the distortion values provide the information only about the magnitude of the error. We still do not know exactly which neighbor's information is lost. To provide this information, we allow users to select a word and show its closest neighbors in the high-dimensional space as a link between the corresponding points. As an additional visual cue, we encode the distance in high-dimensional space as the thickness (thick-close, thin-far) of the line. Finally, we support multiple types of clustering algorithms to allow an interactive exploration of both the high-dimensional neighbors and clusters.

7.2 Analogy Relationships View

In the word embedding space, analogy pairs exhibit interesting algebraic relationships (see Fig. 2) that are often used to evaluate the quality of word embeddings [29]. Consequently, obtaining an in-depth understanding of analogy pairs' behaviors is essential for exploring the word embeddings. The analogy relationships view is designed to address tasks T4-T8, and to some extent T9 in the task list. As illustrated in Fig. 7, there are three panels in the analogy relationship view.

Analogy Pairs Projection. The analogy pair projection panel (Fig.7(b)) supports different linear projections: PCA, SVM+REG and SVM+PCA. The orange and blue colors correspond to the two concepts of a given analogy group. The link connects the words belonging to the same analogy. Furthermore, each projection displays its error relative to the corresponding null hypothesis (top left corner) to indicate the salience of the observed structure. Both SVM-based approaches capture hidden subspaces (T5, T6), which are typically lost in a PCA embedding. Each of the methods emphasizes a different aspect of the data. As illustrated in our case study (Section 8), by examining their relationships, the user can obtain a multifaceted understanding of analogy pairs' structure.

The system also provides an animated transition between any pair of linear projections (e.g., from PCA to SVM+PCA). Compared to showing different projections sequentially, the dynamic transitions [9, 22] allow users to maintain the visual context and track the correspondence between individual words. In addition, each frame in the animation is always a linear projection (i.e., a generalization of 3D rotation in high-dimensional space). As illustrated in the case study, the transition provides the user with an intuitive understanding of how one subspace is related to another in a geometric sense, akin to how a 3D rotation helps convey geometric relationships.

In addition, this panel provides the option of applying a variety of clustering algorithms. It includes not only widely used hierarchical, k-means++ and spectral clustering, but also advanced methods such as subspace clustering [11] that are designed to reveal the low-dimensional subspaces shared by subsets of data. As demonstrated in Section 8.1, subspace clustering is ideal for identifying subtrends and other intricate linear relationships within an analogy group.

Pairwise Cosine Distance Histogram. Linear projections of an analogy relationship inform the user whether the pair directions are coherent in a given (2D) subspace. However, knowing the actual angles between the pair directions in the high-dimensional vector space is also important. As illustrated in Fig. 7(c), a histogram of all pairwise cosine distances in an analogy group conveys quantitatively how coherent the analogy relationship is. Combined with the test statistic of the projections, the distribution of distances represents an intuitive way to judge how confident one should be in interpreting the projections. In the histogram, the horizontal axis corresponds to the cosine distance (0.0-1.0).

Semantic Axis. According to one of the domain experts, researchers in NLP suspect that there exist correlations between certain dimensions (factors) and a specific concept or meaning. As demonstrated in Section 5, we can find the general direction of the analogy from the SVM direction. However, the projection and the histogram panel alone do not readily address global inquiries such as the task T7, where the hypothesis needs to be evaluated in the global word embedding space.

The semantic axis panel is designed to identify which words in the global embedding space have the largest or smallest values along the given axis defined by a vector direction (a factor of the word embedding dimensions), for example, an analogy pair, an analogy group, or a concept simply defined by two words (e.g., *the "royalty":man-king*). If the given axis corresponds to a distinct concept, such as masculine and feminine, one expects to find masculine and feminine words at the opposite ends of the axis, which we refer to as the semantic axis.

As shown in Fig. 7(d), the horizontal axis corresponds to values of the words on the semantic axis, and the vertical axis corresponding to the ranking order based on the same value. Note that we show only the k top- and bottom-most words along an axis to keep the list size manageable and focus on the most important words. Fig. 7(d) shows

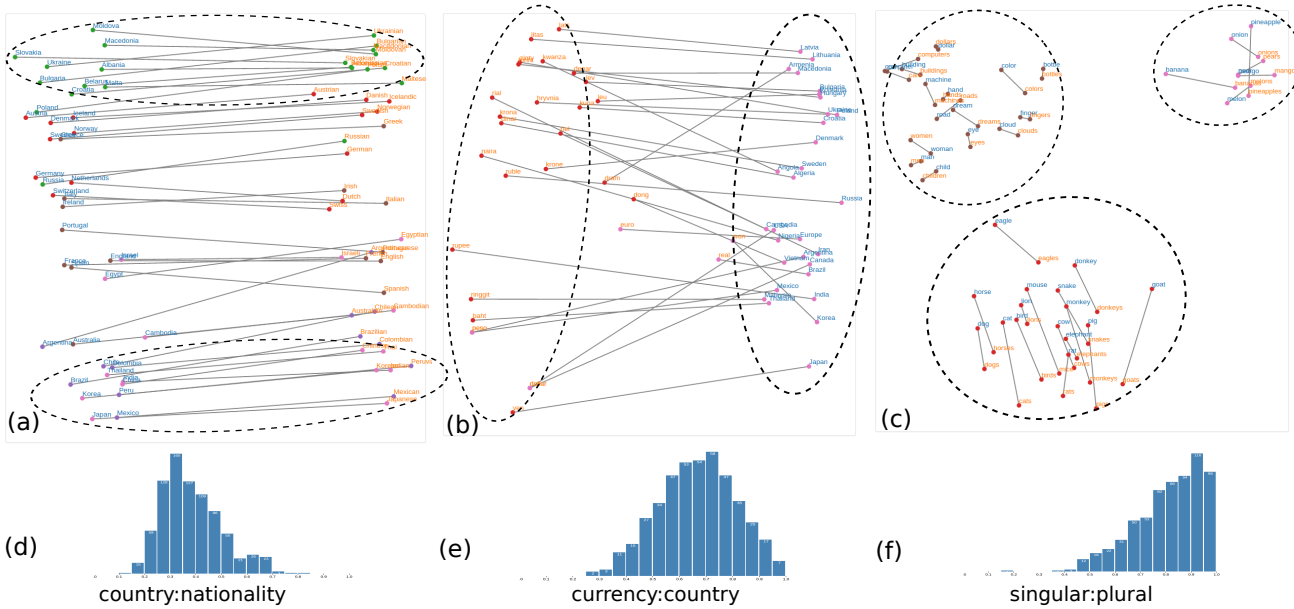


Fig. 8. Different types of analogies can correspond to very different structures. In (a), (b), (c), the subspace clusters (highlighted by dotted circles) are identified in each of the analogy groups. In (a), closely related analogy pairs are grouped into the same cluster. In (b), *currency*:*country*, the two concepts in the analogy are distinct, so the subspace clustering focused on the linear trends within the two groups, instead of among the analogy pairs. In (c), the *singular*:*plural* analogy group contains words that have very different concepts (animal, fruit, etc.); therefore, analogy pairs are grouped into clusters not necessarily because they have a similar orientation, but because the larger distances between these different concepts have a stronger influence. The histograms in (d), (e), (f) confirm the observation.

the 20 top- and bottom-most words along the male-female direction among a list of 10k most frequent words. Conceptually, these words represent what the embedding defines as most masculine vs. feminine. Unsurprisingly, the results show male/female names and gender stereotypical jobs. By examining these words, we can verify the authenticity of the concept obtained using a limited number of words in the global embedding space.

7.3 Implementation

As illustrated in Fig. 9, The system is split into server and client modules. The server handles complex computation tasks and the client manages the user interface. The client is web-based, allowing us to continuously share the latest improvements with our collaborators, which has been crucial for a tight design, implementation, and feedback circle. The communication between the web client and the server is accomplished through a set of RESTful APIs.

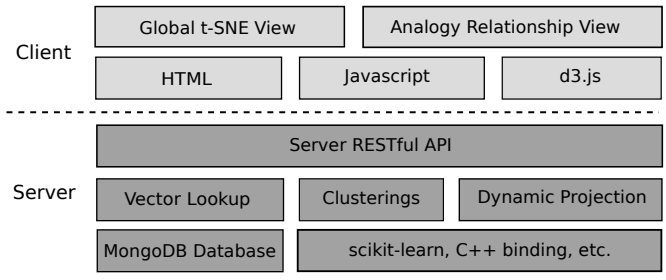


Fig. 9. An overview of the system architecture. The entire system is split into server and client modules, where the server handles complex computation tasks and the client handles user interaction and display.

To achieve a good trade-off between implementation complexity and performance, the server is implemented in *Python*, and the computation methods are handled by efficient libraries (e.g., scikit-learn [32]) or python bindings of native C++ code. The client graphical interface is implemented in *JavaScript* and *HTML* with *d3.js* [4] to handle the graphical components. The widely used pretrained datasets usually contain large numbers of words and phrases (i.e., word2vec's googleNews dataset contains 3 million words and phrases). To manage the pretrained and user-provided data and achieve efficient storage and query operations, we use the MongoDB [1] database on the server.

8 Case Study: Analogy Tasks

In this section, we showcase how the domain scientists have used our tool to study semantic and syntactic relationships in word embeddings and gained new insights. The analogy task dataset (originally used in [27]) examined in this study is one of the most widely used analogy datasets for evaluating the quality of word embeddings. It contains 14 analogy groups, some of which are semantic (*male*:*female*) in nature, whereas the others are syntactic (*singular*:*plural*). For each analogy group, a set of analogy pairs (e.g., *man*:*woman*) is provided. This dataset is often used to compute the error in analogy prediction (i.e., is *man* - *woman* + *queen* close to *king*?), which is considered as a key indicator of how well the word embedding captures the intricate relationships among words. Despite being a common practice, little is known regarding the characteristics of these analogy relationships, and how they compare to each other within the same or different embeddings (e.g., Glove vs. word2vec). The goal of this study is to address the often-neglected and challenging questions (T5-T9) about analogy relationships. In the following visualizations, the 300D version of pretrained word embeddings (Glove or word2vec) is used.

8.1 Are They Really Parallel?

The assumption of analogy relationship (as illustrated in Fig. 2) suggests that the pair orientations (e.g., *man* - *women*) within each analogy group should be similar to each other. Particularly, 2D PCA projections, in which the pair directions are parallel to each other, are typically used to illustrate the coherency of the analogy relationship in the NLP community. However, projections can easily destroy existing correlations between directions (i.e., by projecting along the analogy direction) or create a false alignment. As a result, without additional information, PCA projections may be misleading.

In our tool, we use the histogram of the pairwise analogy directions to provide a direct estimate of the parallelism among the orientations in high-dimensional space. As illustrated in Fig. 8(a),(d) for the *country*:*nationality* analogy, even though in the PCA projection the direc-

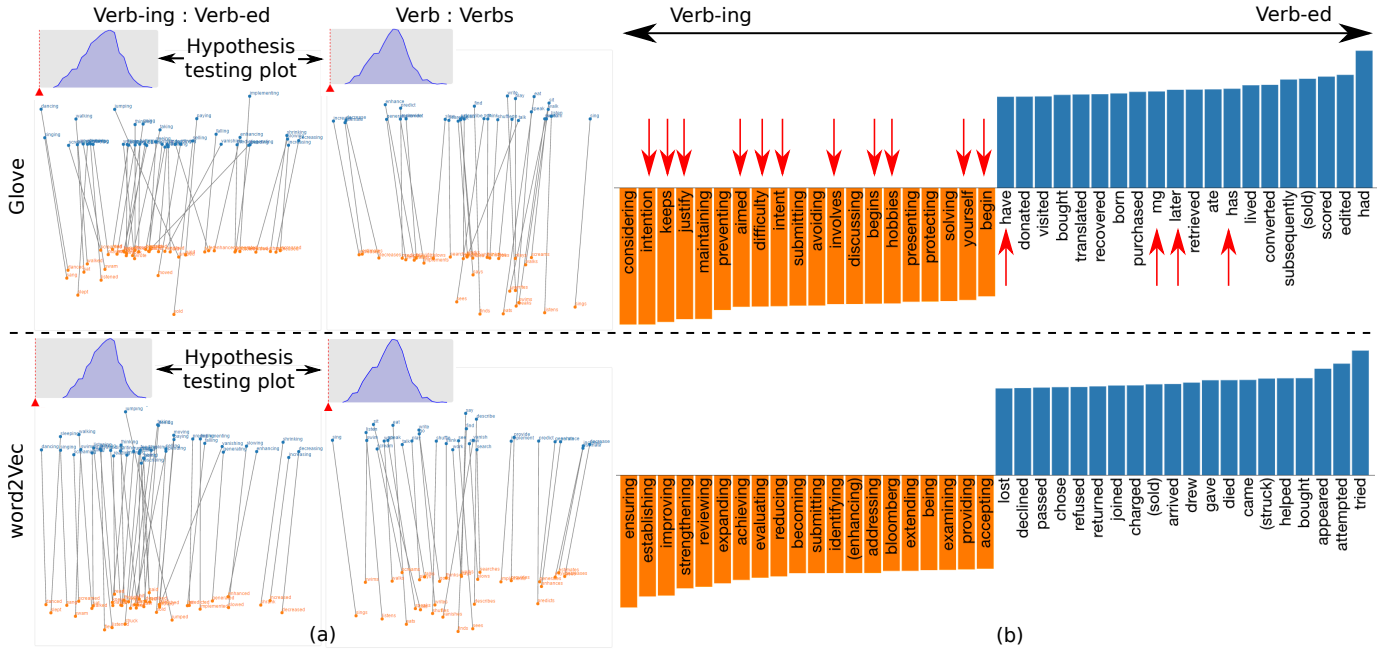


Fig. 10. Comparison of Glove and word2Vec word embeddings: (a) Analogy projection (SVM+PCA)-based comparison shows that word2Vec produces a more apparently aligned directions for both syntactic relationships. Here we utilize the hypothesis-testing plot to verify whether the projection captures the truthful structure of the data. As we can see in all four plots, the current error (red triangle) falls outside the possible range of error from random data, which indicates high confidence in the presence of salience structures. (b) The semantic axis plots reveal that the average analogy direction from word2Vec embedding better identifies the true syntactic concept (*verb-ing*:*verb-ed*) word2Vec embedding. The mistakes in the case of Glove are marked with red arrows.

tions appear similar, the histogram reveals that most pairs are not as well aligned (with cosine distance value 0.3-0.5) as they appear. One of the domain experts was surprised by numerous such examples when exploring the different analogy groups using the tool. It is unexpected because *country*:*nationality* is one of the analogy groups that has the most coherent-looking (parallelism among the lines) PCA embedding. Furthermore, when examining all the 14 analogy pairs, the scientist noticed that different types of analogy groups can have extremely different behaviors (see Fig 8).

To help make sense of these huge variations among the analogy relationships, the scientist could utilize the different projection techniques and subspace clustering [11, 44] (group the data as belonging to different low-dimensional subspaces, discussed in Section 7.2) to shed light on the structure of words in each analogy group. In Fig. 8(a), for the case of *country*:*nationality*, closely related analogy pairs are grouped into the same cluster (e.g., Scandinavian countries and capitals, highlighted by the dotted circle at the top of the embedding). Whereas, as seen in Fig. 8(b), the two concepts in the *currency*:*country* analogy are distinct, so when applying subspace clustering, we see currencies and countries forming their own subspaces. In other words, subspace clustering identifies the stronger linear trends within currencies and countries, instead of analogy pairs. Finally, in Fig. 8(c), the *singular*:*plural* analogy group contains words that have very different concepts (*animal*:*animals* vs. *fruit*:*fruits*), and therefore, when applying subspace clustering, animal and fruit words are grouped into different subspaces, not necessarily because the pairs have very similar directions, but because the differences between concepts have a stronger influence on the subspace distance.

8.2 Do Analogy Directions Really Capture Semantics?

Based on the variations observed among pair orientations in each analogy group, the domain scientist naturally started questioning the widely accepted assumption that an analogy direction corresponds to a particular semantic idea. If the variation is so large within an analogy such as the *singular*:*plural* case, does the analogy direction still encode the actual syntactic relationship?

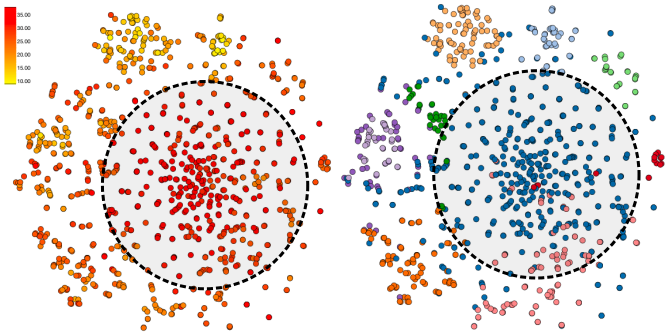
To further investigate this conundrum, the domain scientist could use the proposed analogy projection schemes to emphasize the apparent word relationships. As described earlier, the projection techniques attempt to find a subspace that maximally reveals the semantic and syntactic relationships. As illustrated in Fig. 4, even for the least coherent analogy group (*singular*:*plural*), as determined by PCA, such a subspace exists (Fig. 4(c)). Furthermore, using the new hypothesis-testing procedure, one can be fairly confident in attributing this result to the inherent structure in the high-dimensional space. The interesting results of our analogy projection enabled the domain expert to hypothesize that the word embedding attempts to preserve the analogy relationship in the high-dimensional space, while simultaneously trying to capture other conflicting relationships. By viewing the dynamic projection transition between PCA and the SVM+PCA projection, the domain scientist could easily track how each analogy pair changes. The animation also revealed an interesting rotational movement, which enabled an intuitive understanding of the geometric relationship between the two projections (see the supplemental video for the animation).

Despite its usefulness for exploratory analysis, the analogy projection focuses only on local relationships. The next natural question of the expert was if the semantic/syntactic relationship observed locally will hold in a global context. And more importantly, is there an apparent relationship between a latent factor (a linear combination of original dimensions) and the corresponding semantic or syntactic concepts? The semantic axis panel provides the domain scientist with the capability to answer these questions. As described in the previous sections, the overall analogy direction of an analogy group can be obtained via the normal of the linear SVM. Fig. 10(b) shows the most extreme words among the 10k most frequent ones along the analogy direction (the semantic axis). Interestingly, one finds words that strongly represent the concept defined by the analogy although they do not actually exist in the original analogy group.

The optimized analogy projections and the semantic axis plots can also aid in the comparison of the different word embedding approaches (T9). As illustrated in Fig. 10, word2Vec is doing better at capturing syntactic relationships than Glove, which explains the superiority of

word2Vec in syntactic analogy prediction tasks commonly observed by NLP researchers [33].

For the domain expert, the new insights from the proposed analysis raised additional questions on the validity and possible limitations of how word embedding quality is evaluated. Using just the algebraic vector relationships in high-dimensional space as the quality measure (as suggested in [29]) might result in the large variance being ignored among different kinds of analogies. Furthermore, as a global error measure, the algebraic relationships do not account for the existence of subspaces. In particular, there can be analogies with rather poor global alignment that nevertheless are highly correlated in some linear subspace. Potentially, this should be taken into account when creating and using word embeddings, which might lead to better results overall. Even though critics (e.g., [20]) from the NLP community have recently drawn attention to many potential problems with the analogy-based evaluation approaches for word embeddings, our visualization still provides a unique perspective for the domain experts on this pressing topic.



(a) Per-Point Embedding Errors (b) Spectral Clustering Labels

Fig. 11. Examining the high distortion regions in the t-SNE.

8.3 What Does the Global Relationship Tell Us?

The analogy data include more than 900 unique words. Studying how they are distributed in the word embedding space may shed light on the differences among analogy groups. The per-point distortion error [17] (discussed in Section 7.1) estimates how well high-dimensional local structure is preserved in the 2D embedding. As illustrated in Fig. 11(a), we can see there is a region in the t-SNE embedding, indicated by the dotted circle, with high distortion errors. By checking the corresponding words, this region corresponds to a broad range of verbs, adverbs, and adjectives. We also notice the different types of nouns correspond to more compact groups. To highlight these groups, we show the spectral clustering label computed from high-dimensional word vectors, which separate the semantically distinct words well. However, other techniques (or word labels) that capture semantic separation can be applied here as well. As showed in Fig. 11(b), various type of nouns (clusters distributed around the perimeter of the embedding) seem to form easily separable regions, but it is much harder to distinguish the differences among verbs, adverbs, and adjectives (indicated by the dotted circle).

The domain scientists found these observations very interesting and postulated that this behavior likely indicates how the vectors are trained. Word2vec's vectors, for example, are created by factorizing the co-occurrence statistics matrix between words (discussed in Section 3). Since nouns (animals, cities, etc) co-occur within fewer contexts than verbs and adjectives, their vector representations will be more sparse, which might lead to tighter clusters.

9 Feedback and Evaluation

We redesigned and improved the system over multiple iterations based on constant feedback from domain scientists. In the beginning of the study, we focused on the effective communication of uncertainty in the t-SNE embedding. Through our subsequent discussions, the importance of analogy relationships became more apparent. Feedback and discussions such as these shaped the goal and features of the tool.

For the evaluation of the final version of the tool, we solicited feedback from three domain experts, who are familiar with the capability of the tool via the usage example and their own experiences from the web-based system, on the following inquiries: 1) Does the tool address the questions summarized in the task list? 2) Does the tool make the user more aware of the uncertainty in the visualization? 3) What is the most useful feature of the tool? 4) Does the new projection better capture analogy relationships? 5) Does the hypothesis testing for salient structure aid in the interpretation of the projection? 6) Does the dynamic transition help explain the relationship between projections? 7) What can be improved? 8) What are the related topics you would like to explore?

A summary of the anecdotal evaluation is as follows: The domain experts agreed that the tool provides a number of unique approaches to address domain-specific inquiries that were not possible before. In particular, one of the experts argued that the most interesting part of the tool is that it helps raise questions he never considered previously (as discussed in the case study). Due to familiarity with t-SNE, all three domain experts were easily drawn to the high-dimensional neighborhood lookup and distortion visualization and found them to be very useful for their everyday workflow. One domain expert pointed out that the concept of hypothesis testing and salient structure estimation can be hard to grasp at first. However, with a little experimentation, he found the information it provides to be invaluable for interpreting the embedding result. Further, he believes that the general idea of hypothesis testing can be valuable for other types of visualization. The domain experts concurred that the addition of the error/uncertainty estimation features made them more aware of the potential pitfall of visualization, which is ignored in a numnber of well known NLP publications. One expert even suggested that he will start promoting the concept of "error in visualization" to the NLP community. During our discussions, we realized that the notion of error in the t-SNE projection is often wrongly interpreted by NLP researchers as an error in the word embedding space and not in the visualization, and our tool helped to clarify this difference to all three domain experts. Finally, in regard to the dynamic transition between linear projections, one domain expert commented that the effect is helpful in tracking the changes between projections. However, he found it to be very challenging to interpret the geometry in high-dimensional spaces since our intuition is built around rotations in 3D. On potential improvements and extensions, one domain expert wonders how could the tool be extended to handle other types of embeddings, such as sentence embeddings [31] instead of words.

10 Conclusion and Future Work

Through a long-term collaboration with NLP domain experts, we introduce the first dedicated visualization tool for exploring word embedding spaces. We have developed a myriad of specialized techniques and visualizations, including linear projection techniques for highlighting analogy relationships, a novel concept of hypothesis testing for salient structure, and various visual encodings for domain-specific tasks. By utilizing the new tool, domain experts are able to gain insights and even form new hypotheses pertinent to language modeling. The concept of word embeddings has recently been extended to phrases or even short sentences [31]. For future work, we plan to extend the scope of the tool to handle these new types of embeddings. On a more general note, we plan to further develop this work into an open-source package in the near future to attract more users and enhance the exposure of our tool in the NLP community.

Acknowledgments

This work is supported in part by NSF: CGV: Award:1314896, NSF:IIP Award :1602127 NSF:ACI:award 1649923, DOE/SciDAC DESC0007446, CCMSC DE-NA0002375, and PIPER: ER26142 DE-SC0010498. This material is based upon work supported by the Department of Energy, National Nuclear Security Administration, under Award Number(s) DE-NA0002375.

REFERENCES

[1] Mongodb: cross-platform document-oriented database. `https://www.mongodb.org/`.

[2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, 2016.

[3] D. J. Biau, B. M. Jolles, and R. Porcher. P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research*, 468(3):885–892, 2010.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[5] K. Brodlie, R. A. Osorio, and A. Lopes. A review of uncertainty in data visualization. In *Expanding the frontiers of visual analytics and visualization*, pages 81–109. Springer, 2012.

[6] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008.

[7] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82(398):424–436, 1987.

[8] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77, 2012.

[9] D. Cook, A. Buja, J. Cabrera, and C. Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 4(3):155–172, 1995.

[10] W. Dou and S. Liu. Topic- and time-oriented visual text analysis. *IEEE Computer Graphics and Applications*, 36(4):8–13, 2016.

[11] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797, 2009.

[12] A. Gladkova and A. Drozd. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of 1st Workshop on Evaluating Vector Space Representations for NLP*, 2016.

[13] A. Gorban and A. Zinovyev. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems*, 20(3):219–232, 2010.

[14] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization'90*, pages 361–378, 1990.

[15] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[16] M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, et al. From word embeddings to document distances. In *International Conference on Machine Learning*, volume 15, pages 957–966, 2015.

[17] J. A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431–1443, 2009.

[18] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Proceedings of Advances in neural information processing systems*, pages 2177–2185, 2014.

[19] J. Li, X. Chen, E. Hovy, and D. Jurafsky. Visualizing and understanding neural models in nlp. In *Proceedings of HLT-NAACL*, 2016.

[20] T. Linzen. Issues in evaluating semantic spaces using word analogies. In *Proceedings of 1st Workshop on Evaluating Vector Space Representations for NLP*, 2016.

[21] S. Liu, B. Wang, P.-T. Bremer, and V. Pascucci. Distortion-guided structure-driven interactive exploration of high-dimensional data. *Computer Graphics Forum*, 33(3):101–110, 2014.

[22] S. Liu, B. Wang, J. J. Thiagarajan, P.-T. Bremer, and V. Pascucci. Visual exploration of high-dimensional data through subspace analysis and dynamic projections. *Computer Graphics Forum*, 34(3):271–280, 2015.

[23] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu. Storyflow: Tracking the evolution of stories. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2436–2445, 2013.

[24] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei. Online visual analytics of text streams. *IEEE Transactions on Visualization and Computer Graphics*, 22(11):2451–2466, 2016.

[25] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 142–150, 2011.

[26] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics*, 41:26–42, 2014.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[28] 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*, pages 3111–3119, 2013.

[29] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013.

[30] B. Mokbel, W. Lueks, A. Gisbrecht, and B. Hammer. Visualizing the quality of dimensionality reduction. *Neurocomputing*, 112:109–123, 2013.

[31] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, 24(4):694–707, 2016.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[33] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, volume 14, pages 1532–1543, 2014.

[34] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, 2012.

[35] X. Rong and E. Adar. Visual tools for debugging neural language models. In *Proceedings of ICML Workshop on Visualization for Deep Learning*, 2016.

[36] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, 2016.

[37] A. Smith, J. Chuang, Y. Hu, J. Boyd-Graber, and L. Findlater. Concurrent visualization of relationships between words and topics in topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, page 7982, 2014.

[38] A. Smith, T. Hawes, and M. Myers. Hiérarchie: Interactive visualization for hierarchical topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 71–78, 2014.

[39] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[40] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565, 2014.

[41] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[42] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(85):2579–2605, 2008.

[43] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[44] R. Vidal. A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, 2011.

[45] C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang. Sentiview: Sentiment analysis and visualization for internet popular topics. *IEEE Transactions on Human-Machine Systems*, 43(6):620–630, 2013.

[46] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Empirical Methods in Natural Language Processing*, pages 1393–1398, 2013.