Semantic Structure and Interpretability of Word Embeddings

Lütfi Kerem Şenel, *Student Member, IEEE*, İhsan Utlu, Veysel Yücesoy, *Student Member, IEEE*, Aykut Koç, *Member, IEEE*, and Tolga Çukur, *Senior Member, IEEE*

***Abstract*—Dense word embeddings, which encode meanings of words to low-dimensional vector spaces, have become very popular in natural language processing (NLP) research due to their state-of-the-art performances in many NLP tasks. Word embeddings are substantially successful in capturing semantic relations among words, so a meaningful semantic structure must be present in the respective vector spaces. However, in many cases, this semantic structure is broadly and heterogeneously distributed across the embedding dimensions making interpretation of dimensions a big challenge. In this study, we propose a statistical method to uncover the underlying latent semantic structure in the dense word embeddings. To perform our analysis, we introduce a new dataset (SEMCAT) that contains more than 6500 words semantically grouped under 110 categories. We further propose a method to quantify the interpretability of the word embeddings. The proposed method is a practical alternative to the classical word intrusion test that requires human intervention.**

***Index Terms*—Interpretability, semantic structure, word embeddings.**

I. Introduction

WORDS are the smallest elements of a language with a practical meaning. Researchers from diverse fields including linguistics [1], computer science [2] and statistics [3] have developed models that seek to capture "word meaning" so that these models can accomplish various NLP tasks such as parsing, word sense disambiguation and machine translation. Most of the effort in this field is based on the distributional hypothesis [4] which claims that a word is characterized by the company it keeps [5]. Building on this idea, several vector space models such as well known Latent Semantic Analysis (LSA) [6] and Latent Dirichlet Allocation (LDA) [7] that make use of word distribution statistics have been proposed in distributional semantics. Although these methods have been commonly used in NLP, more recent techniques that generate dense, continuous valued vectors, called *embeddings*, have been receiving increasing interest in NLP research. Approaches that learn embeddings include neural network based predictive methods [2], [8] and count-based matrix-factorization methods [9]. Word embeddings brought about significant performance improvements in many intrinsic NLP tasks such as analogy or semantic textual similarity tasks, as well as downstream NLP tasks such as part-of-speech (POS) tagging [10], named entity recognition [11], word sense disambiguation [12], sentiment analysis [13] and cross-lingual studies [14].

Although high levels of success have been reported in many NLP tasks using word embeddings, the individual embedding dimensions are commonly considered to be uninterpretable [15]. Contrary to some earlier sparse vector space models such as Hyperspace Analogue to Language (HAL) [16], what is represented in each dimension of word embeddings is often unclear, rendering them a black-box approach. In contrast, embedding models that yield dimensions that are more easily interpretable in terms of the captured information can be better suited for NLP tasks that require semantic interpretation, including named entity recognition and retrieval of semantically related words. Model interpretability is also becoming increasingly relevant from a regulatory standpoint, as evidenced by the recent EU regulation that grants people with a "right to explanation" regarding automatic decision making algorithms [17].

Although word embeddings are a dominant part of NLP research, most studies aim to maximize the task performance on standard benchmark tests such as MEN [18] or Simlex-999 [19]. While improved test performance is undoubtedly beneficial, an embedding with enhanced performance does not necessarily reveal any insight about the semantic structure that it captures. A systematic assessment of the semantic structure intrinsic to word embeddings would enable an improved understanding of this popular approach, would allow for comparisons among

Manuscript received November 22, 2017; revised April 12, 2018; accepted May 10, 2018. Date of publication May 24, 2018; date of current version June 21, 2018. This work was supported in part by the European Molecular Biology Organization Installation under Grant IG 3028, in part by the TUBA GEBIP fellowship, and in part by the BAGEP 2017 award of the Science Academy. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Imed Zitouni. T. Çukur and A. Koç mutually supervised this work under a joint industry-university coadvising program. *(Corresponding author: Lütfi Kerem Şenel.)*

L. K. Şenel is with the ASELSAN Research Center, Ankara 06370, Turkey, with the Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey, and also with the UMRAM, Bilkent University, Ankara 06800, Turkey (e-mail: lksenel@aselsan.com.tr).

İ. Utlu and V. Yücesoy are with the ASELSAN Research Center, Ankara 06370, Turkey, and also with the Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: utlu@ee.bilkent.edu.tr; vyucesoy@aselsan.com.tr).

A. Koç is with the ASELSAN Research Center, Ankara 06370, Turkey (e-mail: aykutkoc@aselsan.com.tr).

T. Çukur is with the Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey, with the UMRAM, Bilkent University, Ankara 06800, Turkey, and also with the Neuroscience Program, Sabuncu Brain Research Center, Bilkent University, Ankara 06800, Turkey (e-mail: cukur@ee.bilkent.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2837384

different embeddings in terms of interpretability and potentially motivate new research directions.

In this study, we aim to bring light to the semantic concepts implicitly represented by various dimensions of a word embedding. To explore these hidden semantic structures, we leverage the category theory [20] that defines a category as a grouping of concepts with similar properties. We use human-designed category labels to ensure that our results and interpretations closely reflect human judgements. Human interpretation can make use of any kind of semantic relation among words to form a semantic group (category). This does not only significantly increase the number of possible categories but also makes it difficult and subjective to define a category. Although several lexical databases such as WordNet [1] have a representation for relations among words, they do not provide categories as needed for this study. Since there is no gold standard for semantic word categories to the best of our knowledge, we introduce a new category dataset where more than 6,500 different words are grouped in 110 semantic categories. Then, we propose a method based on distribution statistics of category words within the embedding space in order to uncover the semantic structure of the dense word vectors. We apply quantitative and qualitative tests to substantiate our method. Finally, we claim that the semantic decomposition of the embedding space can be used to quantify the interpretability of the word embeddings without requiring any human effort unlike the word intrusion test [21].

This paper is organized as follows: Following a discussion of related work in Section II, we describe our methods in Section III. In this section we introduce our dataset and also describe methods we used to investigate the semantic decomposition of the embeddings, to validate our findings and to measure the interpretability. In Section IV, we present the results of our experiments and finally we conclude the paper in Section V.

II. Related Work

In the word embedding literature, the problem of interpretability has been approached via several different routes. For learning sparse, interpretable word representations from co-occurrence variant matrices, [22] suggested algorithms based on non-negative matrix factorization (NMF) and the resulting representations are called non-negative sparse embeddings (NNSE). To address memory and scale issues of the algorithms in [22], [23] proposed an online method of learning interpretable word embeddings. In both studies, interpretability was evaluated using a word intrusion test introduced in [21]. The word intrusion test is expensive to apply since it requires manual evaluations by human observers separately for each embedding dimension. As an alternative method to incorporate human judgement, [24] proposed joint non-negative sparse embedding (JNNSE), where the aim is to combine text-based similarity information among words with brain activity based similarity information to improve interpretability. Yet, this approach still requires labor-intensive collection of neuroimaging data from multiple subjects.

Instead of learning interpretable word representations directly from co-occurrence matrices, [25] and [26] proposed to use sparse coding techniques on conventional dense word embeddings to obtain sparse, higher dimensional and more interpretable vector spaces. However, since the projection vectors that are used for the transformation are learned from the word embeddings in an unsupervised manner, they do not have labels describing the corresponding semantic categories. Moreover, these studies did not attempt to enlighten the dense word embedding dimensions, rather they learned new high dimensional sparse vectors that perform well on specific tests such as word similarity and polysemy detection. In [26], interpretability of the obtained vector space was evaluated using the word intrusion test. An alternative approach was proposed in [27], where interpretability was quantified by the degree of clustering around embedding dimensions and orthogonal transformations were examined to increase interpretability while preserving the performance of the embedding. Note, however, that it was shown in [27] that total interpretability of an embedding is constant under any orthogonal transformation and it can only be redistributed across the dimensions. With a similar motivation to [27], [28] proposed rotation algorithms based on exploratory factor analysis (EFA) to preserve the expressive performance of the original word embeddings while improving their interpretability. In [28], interpretability was calculated using a distance ratio (DR) metric that is effectively proportional to the metric used in [27]. Although interpretability evaluations used in [27] and [28] are free of human effort, they do not necessarily reflect human interpretations since they are directly calculated from the embeddings.

Taking a different perspective, a recent study, [29], attempted to elucidate the semantic structure within NNSE space by using categorized words from the HyperLex dataset [30]. The interpretability levels of embedding dimensions were quantified based on the average values of word vectors within categories. However, HyperLex is constructed based on a single type of semantic relation (hypernym) and average number of words representing a category is significantly low ($\approx$2) making it challenging to conduct a comprehensive analysis.

III. Methods

To address the limitations of the approaches discussed in Section II, in this study we introduce a new conceptual category dataset. Based on this dataset, we propose statistical methods to capture the hidden semantic concepts in word embeddings and to measure the interpretability of the embeddings.

A. Dataset

Understanding the hidden semantic structure in dense word embeddings and providing insights on interpretation of their dimensions are the main objectives of this study. Since embeddings are formed via unsupervised learning on unannotated large corpora, some conceptual relationships that humans anticipate may be missed and some that humans do not anticipate may be formed in the embedding space [31]. Thus, not all clusters obtained from a word embedding space will be interpretable. Therefore, using the clusters in the dense embedding space might not take us far towards interpretation. This observation is also rooted in the need for human judgement in evaluating interpretability.

To provide meaningful interpretations for embedding dimensions, we refer to the category theory [20] where concepts with similar semantic properties are grouped under a common category. As mentioned earlier, using clusters from the embedding space as categories may not reflect human expectations accurately, hence having a basis based on human judgements is essential for evaluating interpretability. In that sense, semantic categories as dictated by humans can be considered a gold standard for categorization tasks since they directly reflect human expectations. Therefore, using supervised categories can enable a proper investigation of the word embedding dimensions. In addition, by comparing the human-categorized semantic concepts with the unsupervised word embeddings, one can acquire an understanding of what kind of concepts can or cannot be captured by the current state-of-the-art embedding algorithms.

In the literature, the concept of category is commonly used to indicate super-subordinate (hyperonym-hyponym) relations where words within a category are types or examples of that category. For instance, the furniture category includes words for furniture names such as bed or table. The HyperLex category dataset [30], which was used in [29] to investigate embedding dimensions, is constructed based on this type of relation that is also the most frequently encoded relation among sets of synonymous words in the WordNet database [1]. However, there are many other types of semantic relations such as meronymy (part-whole relations), antonymy (opposite meaning words), synonymy (words having the same sense) and cross-Part of Speech (POS) relations (i.e., lexical entailments). Although WordNet provides representations for a subset of these relations, there is no clear procedure for constructing unified categories based on multiple different types of relations. It remains unclear what should be considered as a category, how many categories there should be, how narrow or broad they should be, and which words they should contain. Furthermore, humans can group words by inference, based on various physical or numerical properties such as color, shape, material, size or speed, increasing the number of possible groups almost unboundedly. For instance, words that may not be related according to classical hypernym or synonym relations might still be grouped under a category due to shared physical properties: sun, lemon and honey are similar in terms of color; spaghetti, limousine and sky-scanner are considered as tall; snail, tractor and tortoise are slow.

In sum, diverse types of semantic relationships or properties can be leveraged by humans for semantic interpretation. Therefore, to investigate the semantic structure of the word embedding space using categorized words, we need categories that represent a broad variety of distinct concepts and distinct types of relations. To the best of our knowledge, there is no comprehensive word category dataset that captures the many diverse types of relations mentioned above. What we have found closest to the required dataset are the online categorized word-lists[1] that were constructed for educational purposes. There are a total of 168 categories on these word-lists. To build a word-category dataset suited for assessing the semantic structure in word embeddings, we took these word-lists as a foundational basis. We filtered out words that are not semantically related but share a common nuisance property such as their POS tagging (verbs, adverbs, adjectives etc.) or being compound words. Several categories containing proper words or word phrases such as the chinese new year and good luck symbols categories, which we consider too specific, are also removed from the dataset. Vocabulary is limited to the most frequent 50,000 words, where frequencies are calculated from English Wikipedia, and words that are not contained in this vocabulary are removed from the dataset. We call the resulting semantically grouped word dataset "SEMCAT[2]" (SEMantic CATegories). Summary statistics of SEMCAT and HyperLex datasets are given in Table I. 10 sample words from each of 6 representative SEMCAT categories are given in Table II.

TABLE I
SUMMARY STATISTICS OF SEMCAT AND HYPERLEX

	SEMCAT	HperLex
Number of Categories	110	1399
Number of Unique Words	6559	1752
Average Word Count per Category	91	2
Standard Deviation of Word Counts	56	3

TABLE II
TEN SAMPLE WORDS FROM EACH OF THE SIX REPRESENTATIVE SEMCAT CATEGORIES

Science	**Sciences**	**Art**	**Car**	**Cooking**	**Geography**
atom	astronomy	abstract	auto	bake	africa
cell	botany	artist	car	barbeque	border
chemical	economics	brush	convertible	boil	capital
data	genetics	composition	hybrid	dough	cartography
element	linguistics	draw	jeep	grill	continent
evolution	neuroscience	masterpiece	limo	juice	earth
laboratory	psychology	photograph	runabout	marinate	east
microscope	taxonomy	perspective	rv	oil	gps
scientist	thermodynamics	sketch	taxi	roast	river
theory	zoology	style	van	serve	sea

B. Semantic Decomposition

In this study, we use GloVe [9] as the source algorithm for learning dense word vectors. The entire content of English Wikipedia is utilized as the corpus. In the preprocessing step, all non-alphabetic characters (punctuations, digits, etc.) are removed from the corpus and all letters are converted to lowercase. Letters coming after apostrophes are taken as separate words (`she'll` becomes `she ll`). The resulting corpus is input to the GloVe algorithm. Window size is set to 15, vector length is chosen to be 300 and minimum occurrence count is set to 20 for the words in the corpus. Default values are used for the remaining parameters. The word embedding matrix, $\mathcal{E}$, is obtained from GloVe after limiting vocabulary to the most frequent 50,000 words in the corpus (i.e. $\mathcal{E}$ is 50,000 $\times$ 300). The GloVe algorithm is again used for the second time on the same corpus generating a second embedding space, $\mathcal{E}^2$, to examine

[1] www.enchantedlearning.com/wordlist/

[2] github.com/avaapm/SEMCATdataset2018

the effects of different initializations of the word vectors prior to training.

To quantify the significance of word embedding dimensions for a given semantic category, one should first understand how a semantic concept can be captured by a dimension, and then find a suitable metric to measure it. [29] assumed that a dimension represents a semantic category if the average value of the category words for that dimension is above an empirical threshold, and therefore took that average value as the representational power of the dimension for the category. Although this approach may be convenient for NNSE, directly using the average values of category words is not suitable for well-known dense word embeddings due to several reasons. First, in dense embeddings it is possible to encode in both positive and negative directions of the dimensions making a single threshold insufficient. In addition, different embedding dimensions may have different statistical characteristics. For instance, average value of the words from the jobs category of SEMCAT is around 0.38 and 0.44 in $221\mathrm{st}$ and 57th dimensions of $\mathcal{E}$ respectively; and the average values across all vocabulary are around 0.37 and -0.05 respectively for the two dimensions. Therefore, the average value of 0.38 for the jobs category may not represent any encoding in the $221\mathrm{st}$ dimension since it is very close to the average of any random set of words in that dimension. In contrast, an average of similar value 0.44 for the jobs category may be highly significant for the 57th dimension. Note that focusing solely on average values might be insufficient to measure the encoding strength of a dimension for a semantic category. For instance, words from the car category have an average of -0.08 that is close to the average across all vocabulary, -0.04, for the 133th embedding dimension. However, standard deviation of the words within the car category is 0.15 which is significantly lower than the standard deviation of all vocabulary, 0.35, for this particular dimension. In other words, although average of words from the car category is very close to the overall mean, category words are more tightly grouped compared to other vocabulary words in the 133th embedding dimension, potentially implying significant encoding.

From a statistical perspective, the question of "How strong a particular concept is encoded in an embedding dimension?" can be interpreted as "How much information can be extracted from a word embedding dimension regarding a particular concept?". If the words representing a concept (i.e. words in a SEMCAT category) are sampled from the same distribution with all vocabulary words, then the answer would be zero since the category would be statistically equivalent to a random selection of words. For dimension i and category j, if $\mathcal{P}_{i,j}$ denotes the distribution from which words of that category are sampled and $\mathcal{Q}_{i,j}$ denotes the distribution from which all other vocabulary words are sampled, then the distance between distributions $\mathcal{P}_{i,j}$ and $\mathcal{Q}_{i,j}$ will be proportional to the information that can be extracted from dimension i regarding category j. Based on this argument, Bhattacharya distance [32] with normal distribution assumption is a suitable metric, which is given in (1), to quantify the level of encoding in the word embedding dimensions. Normality of the embedding dimensions are tested using one-sample Kolmogorov-Smirnov test (KS test, Bonferroni corrected for multiple comparisons).

$$\mathcal{W}_B(i,j) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\sigma^2_{p_{i,j}}}{\sigma^2_{q_{i,j}}} + \frac{\sigma^2_{q_{i,j}}}{\sigma^2_{p_{i,j}}} + 2\right)\right) + \frac{1}{4}\left(\frac{\left(\mu_{p_{i,j}} - \mu_{q_{i,j}}\right)^2}{\sigma^2_{p_{i,j}} + \sigma^2_{q_{i,j}}}\right) \tag{1}$$

In (1), $\mathcal{W}_B$ is a 300×110 Bhattacharya distance matrix, which can also be considered as a category weight matrix, i is the dimension index ($i \in \{1, 2, \ldots, 300\}$), j is the category index ($j \in \{1, 2, \ldots, 110\}$). $p_{i,j}$ is the vector of the ith dimension of each word in jth category and $q_{i,j}$ is the vector of the ith dimension of all other vocabulary words ($p_{i,j}$ is of length n_j and $q_{i,j}$ is of length $(50000 - n_j)$ where n_j is the number of words in the jth category). μ and σ are the mean and the standard deviation operations, respectively. Values in $\mathcal{W}_B$ can range from 0 (if $p_{i,j}$ and $q_{i,j}$ have the same means and variances) to ∞. In general, a better separation of category words from remaining vocabulary words in a dimension results in larger $\mathcal{W}_B$ elements for the corresponding dimension.

Based on SEMCAT categories, for the learned embedding matrices $\mathcal{E}$ and $\mathcal{E}^2$, the category weight matrices ($\mathcal{W}_B$ and $\mathcal{W}_B^2$) are calculated using Bhattacharya distance metric (1).

C. Interpretable Word Vector Generation

If the weights in $\mathcal{W}_B$ truly correspond to the categorical decomposition of the semantic concepts in the dense embedding space, then $\mathcal{W}_B$ can also be considered as a transformation matrix that can be used to map word embeddings to a semantic space where each dimension is a semantic category. However, it would be erroneous to directly multiply the word embeddings with category weights. The following steps should be performed in order to map word embeddings to a semantic space where dimensions are interpretable:

1) To make word embeddings compatible in scale with the category weights, word embedding dimensions are standardized ($\mathcal{E}_S$) such that each dimension has zero mean and unit variance since category weights have been calculated based on the deviations from the general mean (second term in (1)) and standard deviations (first term in (1)).
2) Category weights are normalized across dimensions such that each category has a total weight of 1 ($\mathcal{W}_{NB}$). This is necessary since some columns of $\mathcal{W}_B$ dominate others in terms of representation strength (will be discussed in Section IV in more detail). This inequality across semantic categories can cause an undesired bias towards categories with larger total weights in the new vector space. ℓ_1 normalization of the category weights across dimensions is performed to prevent bias.
3) Word embedding dimensions can encode semantic categories in both positive and negative directions ($\mu_{p_{i,j}} - \mu_{q_{i,j}}$ can be positive or negative) that contribute equally to the Bhattacharya distance. However, since encoding directions are important for the mapping of the word embeddings, $\mathcal{W}_{NB}$ is replaced with its signed version $\mathcal{W}_{NSB}$ (if $\mu_{p_{i,j}} - \mu_{q_{i,j}}$ is negative, then

$\mathcal{W}_{NSB}(i,j) = -\mathcal{W}_{NB}(i,j)$, otherwise $\mathcal{W}_{NSB}(i,j) = \mathcal{W}_{NB}(i,j)$) where negative weights correspond to encoding in the negative direction.

Then, interpretable semantic vectors ($\mathcal{I}_{50000\times 110}$) are obtained by multiplying $\mathcal{E}_S$ with $\mathcal{W}_{NSB}$.

One can reasonably suggest to alternatively use the centers of the vectors of the category words as the weights for the corresponding category as given in (2).

$$\mathcal{W}_C(i,j) = \mu_{p_{i,j}} \tag{2}$$

A second interpretable embedding space, $\mathcal{I}^*$, is then obtained by simply projecting the word vectors in $\mathcal{E}$ to the category centers. (3) and (4) show the calculation of $\mathcal{I}$ and $\mathcal{I}^*$ respectively. Fig. 1 shows the procedure for generation of interpretable embedding spaces $\mathcal{I}$ and $\mathcal{I}^*$.

$$\mathcal{I} = \mathcal{E}_S \mathcal{W}_{NSB} \tag{3}$$

$$\mathcal{I}^* = \mathcal{E}\mathcal{W}_C \tag{4}$$

D. Validation

$\mathcal{I}$ and $\mathcal{I}^*$ are further investigated via qualitative and quantitative approaches in order to confirm that $\mathcal{W}_B$ is a reasonable semantic decomposition of the dense word embedding dimensions, that $\mathcal{I}$ is indeed an interpretable semantic space and that our proposed method produces better representations for the categories than their center vectors.

If $\mathcal{W}_B$ and $\mathcal{W}_C$ represent the semantic distribution of the word embedding dimensions, then columns of $\mathcal{I}$ and $\mathcal{I}^*$ should correspond to semantic categories. Therefore, each word vector in $\mathcal{I}$ and $\mathcal{I}^*$ should represent the semantic decomposition of the respective word in terms of the SEMCAT categories. To test this prediction, word vectors from the two semantic spaces ($\mathcal{I}$ and $\mathcal{I}^*$) are qualitatively investigated.

To compare $\mathcal{I}$ and $\mathcal{I}^*$, we also define a quantitative test that aims to measure how well the category weights represent the corresponding categories. Since weights are calculated directly using word vectors, it is natural to expect that words should have high values in dimensions that correspond to the categories they belong to. However, using words that are included in the categories for investigating the performance of the calculated weights is similar to using training accuracy to evaluate model performance in machine learning. Using validation accuracy is more adequate to see how well the model generalizes to new, unseen data that, in our case, correspond to words that do not belong to any category. During validation, we randomly select 60% of the words for training and use the remaining 40% for testing for each category. From the training words we obtain the weight matrix $\mathcal{W}_B$ using Bhattacharya distance and the weight matrix $\mathcal{W}_C$ using the category centers. We select the largest k weights ($k \in \{5, 7, 10, 15, 25, 50, 100, 200, 300\}$) for each category (i.e. largest k elements for each column of $\mathcal{W}_B$ and $\mathcal{W}_C$) and replace the other weights with 0 that results in sparse category weight matrices ($\mathcal{W}_B^s$ and $\mathcal{W}_C^s$). Then projecting dense word vectors onto the sparse weights from $\mathcal{W}_B^s$ and $\mathcal{W}_C^s$, we obtain interpretable semantic spaces $\mathcal{I}_k$ and $\mathcal{I}_k^*$. Afterwards, for each category, we calculate the percentages of the unseen test words that are among the top n, 3n and 5n words (excluding the training words) in their corresponding dimensions in the new spaces, where n is the number of test words that varies across categories. We calculate the final accuracy as the weighted average of the accuracies across the dimensions in the new spaces, where the weighting is proportional to the number of test words within the categories. We repeat the same procedure for 10 independent random selections of the training words.

E. Measuring Interpretability

In addition to investigating the semantic distribution in the embedding space, a word category dataset can be also used to quantify the interpretability of the word embeddings. In several studies, [21]–[23], interpretability is evaluated using the word intrusion test. In the word intrusion test, for each embedding dimension, a word set is generated including the top 5 words in the top ranks and a noisy word (intruder) in the bottom ranks of that dimension. The intruder is selected such that it is in the top ranks of a separate dimension. Then, human editors are asked to determine the intruder word within the generated set. The editors' performances are used to quantify the interpretability of the embedding. Although evaluating interpretability based on human judgements is an effective approach, word intrusion is an expensive method since it requires human effort for each evaluation. Furthermore, the word intrusion test does not quantify the interpretability levels of the embedding dimensions, instead it yields a binary decision as to whether a dimension is interpretable or not. However, using continuous values is more adequate than making binary evaluations since interpretability levels may vary gradually across dimensions.

We propose a framework that addresses both of these issues by providing automated, continuous valued evaluations of interpretability while keeping the basis of the evaluations as human judgements. The basic idea behind our framework is that humans interpret dimensions by trying to group the most distinctive words in the dimensions (i.e. top or bottom rank words), an idea also leveraged by the word intrusion test. Based on this key idea, it can be noted that if a dataset represents all the possible groups humans can form, instead of relying on human evaluations, one can simply check whether the distinctive words of the embedding dimensions are present together in any of these groups. As discussed earlier, the number of groups humans can form is theoretically unbounded, therefore it is not possible to compile an all-comprehensive dataset for all potential groups. However, we claim that a dataset with a sufficiently large number of categories can still provide a good approximation to human judgements. Based on this argument, we propose a simple method to quantify the interpretability of the embedding dimensions.

We define two interpretability scores for an embedding dimension-category pair as:

$$IS_{i,j}^{+} = \frac{|S_j \cap V_i^{+}(\lambda \times n_j)|}{n_j} \times 100$$
$$IS_{i,j}^{-} = \frac{|S_j \cap V_i^{-}(\lambda \times n_j)|}{n_j} \times 100 \tag{5}$$

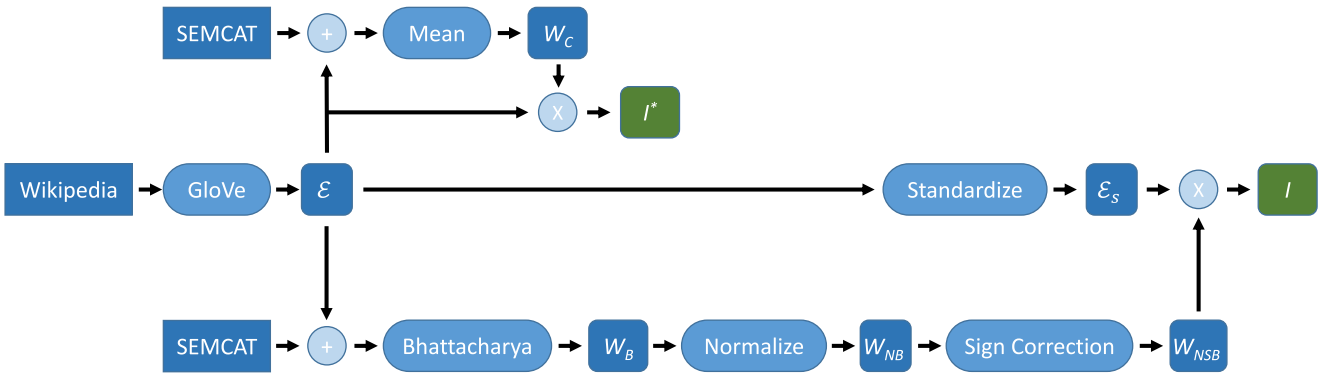

Fig. 1. Flow chart for the generation of the interpretable embedding spaces $\mathcal{I}$ and $\mathcal{I}^*$. First, word vectors are obtained using the GloVe algorithm on Wikipedia corpus. To obtain $\mathcal{I}^*$, weight matrix $\mathcal{W}_C$ is generated by calculating the means of the words from each category for each embedding dimension and then $\mathcal{W}_C$ is multiplied by the embedding matrix (see Section III-C). To obtain $\mathcal{I}$, weight matrix $\mathcal{W}_B$ is generated by calculating the Bhattacharya distance between category words and remaining vocabulary for each category and dimension. Then, $\mathcal{W}_B$ is normalized (see Section III-C item 2), sign corrected (see Section III-C item 3), and finally multiplied with standardized word embedding ($\mathcal{E}_s$, see Section III-C item 1).

where $IS^+_{i,j}$ is the interpretability score for the positive direction and $IS^-_{i,j}$ is the interpretability score for the negative direction for the ith dimension ($i \in \{1, 2, \ldots, D\}$ where D is the dimensionality of the embedding) and jth category ($j \in \{1, 2, \ldots, K\}$ where K is the number of categories in the dataset). S_j is the set representing the words in the jth category, n_j is the number of the words in the jth category and $V^+_i(\lambda \times n_j)$, $V^-_i(\lambda \times n_j)$ refer to the distinctive words located at the top and bottom ranks of the ith embedding dimension, respectively. $\lambda \times n_j$ is the number of words taken from the upper and bottom ranks where λ is the parameter determining how strict the interpretability definition is. The smallest value for λ is 1 that corresponds to the most strict definition and larger λ values relax the definition by increasing the range for selected category words. $\cap$ is the intersection operator between category words and top and bottom ranks words, $|.|$ is the cardinality operator (number of elements) for the intersecting set.

We take the maximum of scores in the positive and negative directions as the overall interpretability score for a category ($IS_{i,j}$). The interpretability score of a dimension is then taken as the maximum of individual category interpretability scores across that dimension (IS_i). Finally, we calculate the overall interpretability score of the embedding (IS) as the average of the dimension interpretability scores:

$$\begin{aligned} IS_{i,j} &= \max(IS^+_{i,j}, IS^-_{i,j}) \\ IS_i &= \max_j IS_{i,j} \\ IS &= \frac{1}{D}\sum_{i=1}^{D} IS_i \end{aligned} \tag{6}$$

We test our method on the GloVe embedding space, on the semantic spaces $\mathcal{I}$ and $\mathcal{I}^*$, and on a random space where word vectors are generated by randomly sampling from a zero mean, unit variance normal distribution. Interpretability scores for the random space are taken as our baseline. We measure the interpretability scores as λ values are varied from 1 (strict interpretability) to 10 (relaxed interpretability).

Our interpretability measurements are based on our proposed dataset SEMCAT, which was designed to be a comprehensive dataset that contains a diverse set of word categories. Yet, it is possible that the precise interpretability scores that are measured here are biased by the dataset used. In general, two main properties of the dataset can affect the results: category selection and within-category word selection. To examine the effects of these properties on interpretability evaluations, we create alternative datasets by varying both category selection and word selection for SEMCAT. Since SEMCAT is comprehensive in terms of the words it contains for the categories, these datasets are created by subsampling the categories and words included in SEMCAT. Since random sampling of words within a category may perturb the capacity of the dataset in reflecting human judgement, we subsample r% of the words that are closest to category centers within each category, where $r \in \{40, 60, 80, 100\}$. To examine the importance of number of categories in the dataset we randomly select m categories from SEMCAT where $m \in \{30, 50, 70, 90, 110\}$. We repeat the selection 10 times independently for each m.

IV. Results

A. Semantic Decomposition

The KS test for normality reveals that 255 dimensions of $\mathcal{E}$ are normally distributed ($p > 0.05$). The average test statistic for these 255 dimensions is 0.0064 ± 0.0016 (mean $\pm$ standard deviation). While the normality hypothesis was rejected for the remaining 45 dimensions, a relatively small test statistic of 0.0156 ± 0.0168 is measured, indicating that the distribution of these dimensions is approximately normal.

The semantic category weights calculated using the method introduced in Section III-B is displayed in Fig. 2. A close examination of the distribution of category weights indicates that the representation of semantic concepts are broadly distributed across many dimensions of the GloVe embedding space. This suggests that the raw space output by the GloVe algorithm has poor interpretability.

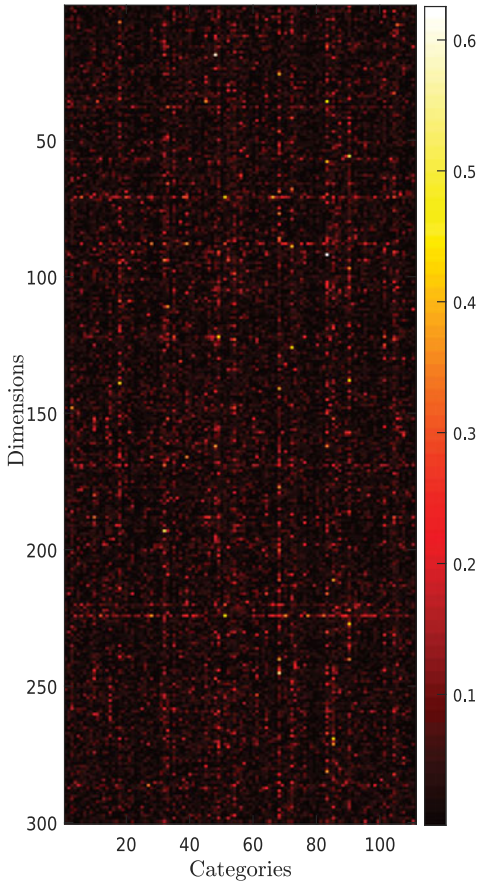

Fig. 2. Semantic category weights ($\mathcal{W}_{B\,300\times110}$) for 110 categories and 300 embedding dimensions obtained using Bhattacharya distance. Weights vary between 0 (represented by black) and 0.63 (represented by white). It can be noticed that some dimensions represent larger number of categories than others do and also some categories are represented strongly by more dimensions than others.

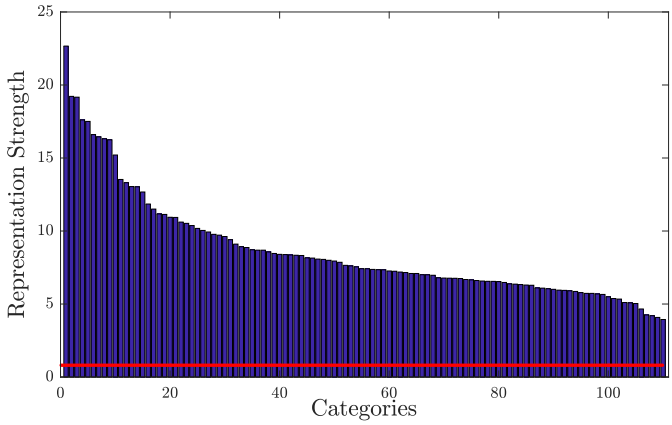

Fig. 3. Total representation strengths of 110 semantic categories from SEMCAT. Bhattacharya distance scores are summed across dimensions and then sorted. Red horizontal line represents the baseline strength level obtained for a category composed of 91 randomly selected words from the vocabulary (where 91 is the average word count across categories in SEMCAT). The metals category has the strongest total representation among SEMCAT categories due to relatively few and well clustered words it contains, while the pirate category has the lowest total representation due to widespread words it contains.

In addition, it can be observed that the total representation strength summed across dimensions varies significantly across categories, some columns in the category weight matrix contain much higher values than others. In fact, total representation strength of a category greatly depends on its word distribution. If a particular category reflects a highly specific semantic concept with relatively few words such as the metals category, category words tend to be well clustered in the embedding space. This tight grouping of category words results in large Bhattacharya distances in most dimensions indicating stronger representation of the category. On the other hand, if words from a semantic category are weakly related, it is more difficult for the word embedding to encode their relations. In this case, word vectors are relatively more widespread in the embedding space, and this leads to smaller Bhattacharya distances indicating that the semantic category does not have a strong representation across embedding dimensions. The total representation strengths of the 110 semantic categories in SEMCAT are shown in Fig. 3, along with the baseline strength level obtained for a category composed of 91 randomly selected words where 91 is the average word count across categories in SEMCAT. The metals category has the strongest total representation among SEMCAT categories due to relatively few and well clustered words it contains, whereas the pirate category has the lowest total representation due to widespread words it contains.

To closely inspect the semantic structure of dimensions and categories, let us investigate the decompositions of three sample dimensions and three specific semantic categories (math, animal and tools). The left column of Fig. 4 displays the categorical decomposition of the 2nd, 6th and 45th dimensions of the word embedding. While the 2nd dimension selectively represents a particular category (sciences), the 45th dimension focuses on 3 different categories (housing, rooms and sciences) and the 6th dimension has a distributed and relatively uniform representation of many different categories. These distinct distributional properties can also be observed in terms of categories as shown in the right column of Fig. 4. While only few dimensions are dominant for representing the math category, semantic encodings of the tools and animals categories are distributed across many embedding dimensions.

Note that these results are valid regardless of the random initialization of the GloVe algorithm while learning the embedding space. For the weights calculated for our second GloVe embedding space $\mathcal{E}^2$, where the only difference between $\mathcal{E}$ and $\mathcal{E}^2$ is the independent random initializations of the word vectors before training, we observe nearly identical decompositions for the categories ignoring the order of the dimensions (similar number of peaks and similar total representation strength; not shown).

B. Validation

A representative investigation of the semantic space $\mathcal{I}$ is presented in Fig. 5, where semantic decompositions of 4 different words, *window*, *bus*, *soldier* and *article*, are displayed using 20 dimensions of $\mathcal{I}$ with largest values for each word. These words are expected to have high values in the dimensions that encode

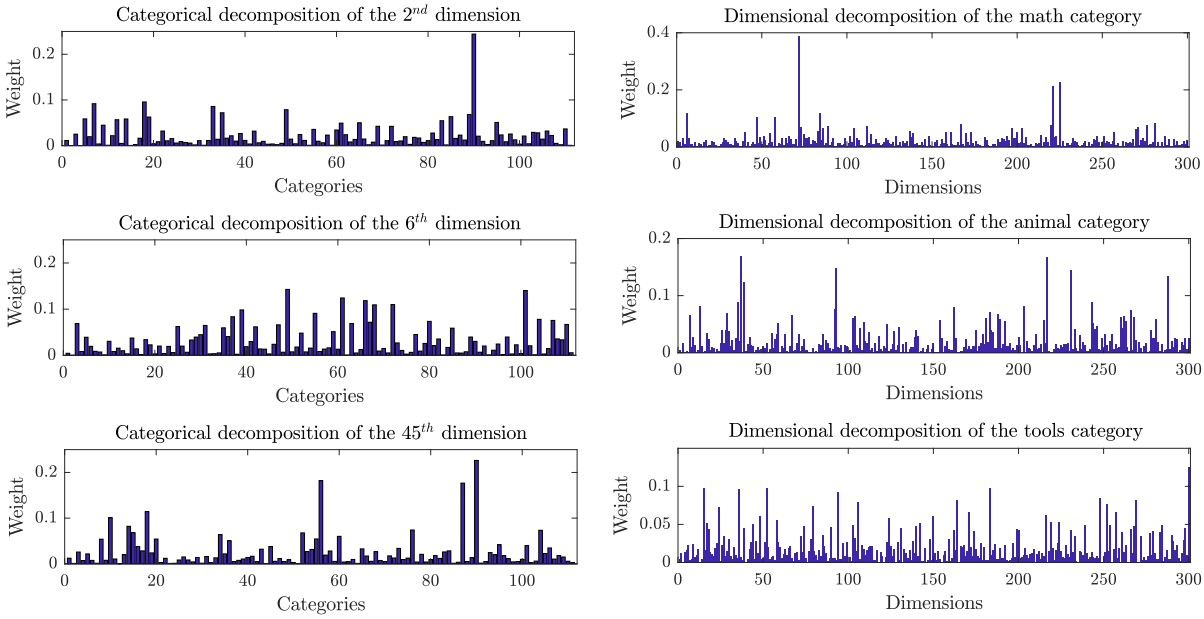

Fig. 4. Categorical decompositions of the 2nd, 6th, and 45th word embedding dimensions are given in the left column. A dense word embedding dimension may focus on a single category (top row), may represent a few different categories (bottom row) or may represent many different categories with low strength (middle row). Dimensional decompositions of the math, animal, and tools categories are shown in the right column. Semantic information about a category may be encoded in a few word embedding dimensions (top row) or it can be distributed across many of the dimensions (bottom row).

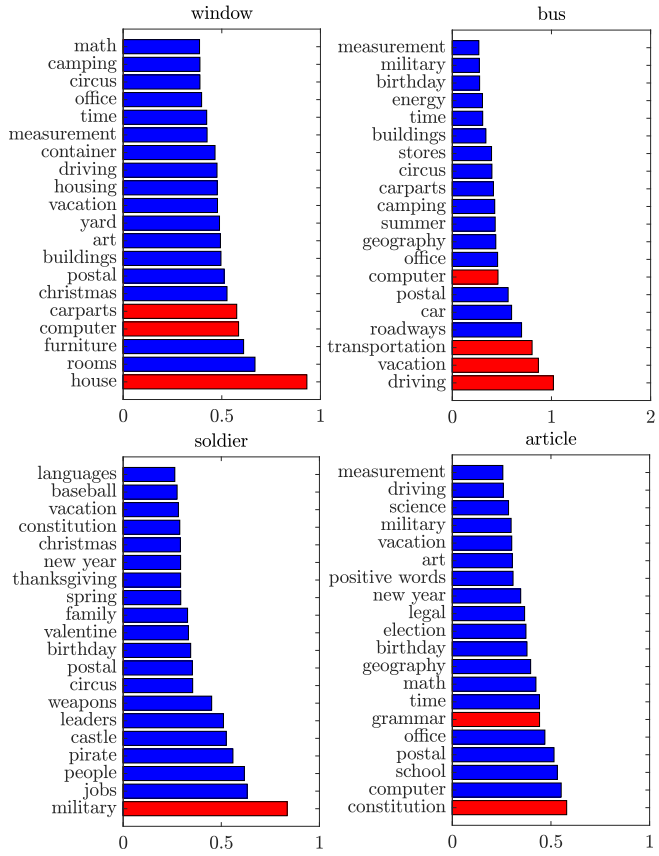

Fig. 5. Semantic decompositions of the words *window*, *bus*, *soldier,* and *article* for 20 highest scoring SEMCAT categories obtained from vectors in $\mathcal{I}$. Red bars indicate the categories that contain the word, blue bars indicate the categories that do not contain the word.

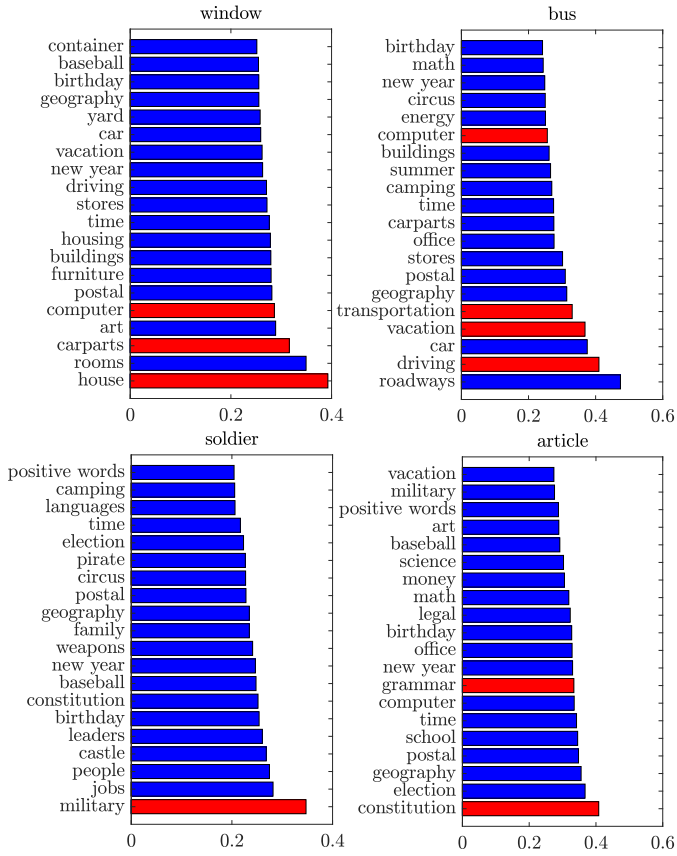

Fig. 6. Categorical decompositions of the words *window*, *bus*, *soldier,* and *article* for 20 highest scoring categories obtained from vectors in $\mathcal{I}^*$. Red bars indicate the categories that contain the word, blue bars indicate the categories that do not contain the word.

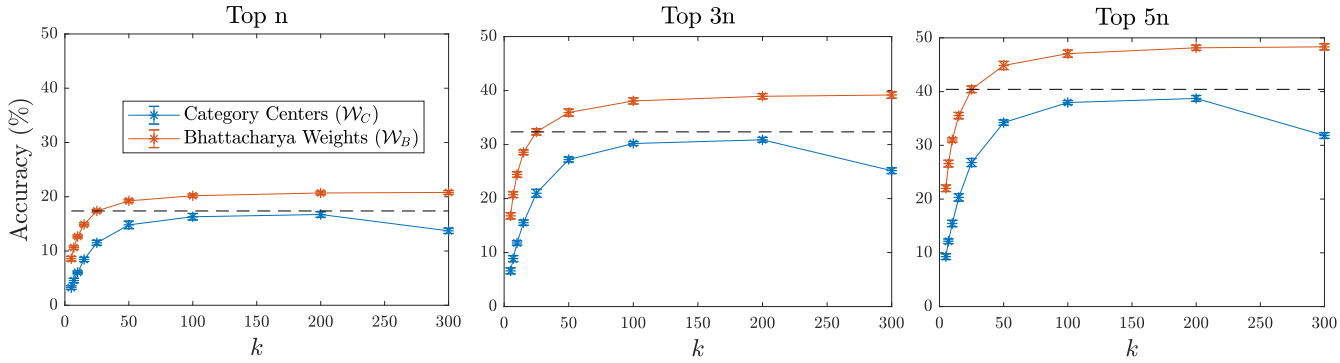

Fig. 7. Category word retrieval performances for top n, $3n$, and $5n$ words where n is the number of test words varying across categories. Category weights obtained using Bhattacharya distance represent categories better than the center of the category words. Using only 25 largest weights from $\mathcal{W}_B$ for each category ($k = 25$) gives better performance than using category centers with any k (shown with dashed line).

the categories to which they belong. However, we can clearly see from Fig. 5 that additional categories such as jobs, people, pirate and weapons that are semantically related to soldier but that do not contain the word also have high values. Similar observations can be made for *window*, *bus*, and *article* supporting the conclusion that the category weight spread broadly to many non-category words.

Fig. 6 presents the semantic decompositions of the words window, bus, soldier and article obtained form $\mathcal{I}^*$ that is calculated using the category centers. Similar to the distributions obtained in $\mathcal{I}$, words have high values for semantically-related categories even when these categories do not contain the words. In contrast to $\mathcal{I}$, however, scores for words are much more uniformly distributed across categories, implying that this alternative approach is less discriminative for categories than the proposed method.

To quantitatively compare $\mathcal{I}$ and $\mathcal{I}^*$, category word retrieval test is applied and the results are presented in Fig. 7. As depicted in Fig. 7, the weights calculated using our method ($\mathcal{W}_B$) significantly outperform the weights from the category centers ($\mathcal{W}_C$). It can be noticed that, using only 25 largest weights from $\mathcal{W}_B$ for each category ($k = 25$) yields higher accuracy in word retrieval compared to the alternative $\mathcal{W}_C$ with any k. This result confirms the prediction that the vectors that we obtain for each category (i.e. columns of $\mathcal{W}_B$) distinguish categories better than their average vectors (i.e. columns of $\mathcal{W}_C$).

C. Measuring Interpretability

Fig. 8 displays the interpretability scores of the GloVe embedding, $\mathcal{I}, \mathcal{I}^*$ and the random embedding for varying λ values. λ can be considered as a design parameter adjusted according to the interpretability definition. Increasing λ relaxes the interpretability definition by allowing category words to be distributed on a wider range around the top ranks of a dimension. We propose that $\lambda = 5$ is an adequate choice that yields a similar evaluation to measuring the top-5 error in category word retrieval tests. As clearly depicted, semantic space $\mathcal{I}$ is significantly more interpretable than the GloVe embedding as justified in Section IV-B. We can also see that interpretability score of

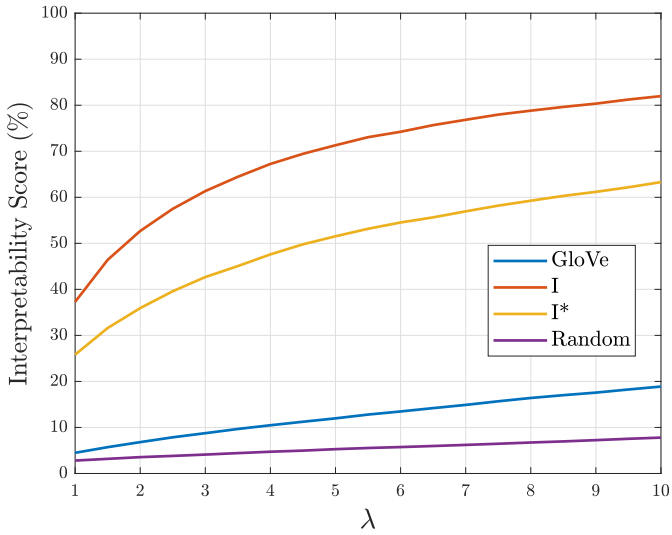

Fig. 8. Interpretability scores for GloVe, $\mathcal{I}$, $\mathcal{I}^*$ and random embeddings for varying λ values where λ is the parameter determining how strict the interpretability definition is ($\lambda = 1$ is the most strict definition, $\lambda = 10$ is a relaxed definition). Semantic spaces $\mathcal{I}$ and $\mathcal{I}^*$ are significantly more interpretable than GloVe as expected. $\mathcal{I}$ outperforms $\mathcal{I}^*$ suggesting that weights calculated with our proposed method more distinctively represent categories as opposed weights calculated as the category centers. Interpretability scores of Glove are close to the baseline (Random) implying that the dense word embedding has poor interpretability.

the GloVe embedding is close to the random embedding representing the baseline interpretability level.

Interpretability scores for datasets constructed by subsampling SEMCAT are given in Table III for the GloVe, $\mathcal{I}$, $\mathcal{I}^*$ and random embedding spaces for $\lambda = 5$. Interpretability scores for all embeddings increase as the number of categories in the dataset increase (30, 50, 70, 90, 110) for each category coverage (40%, 60%, 80%, 100%). This is expected since increasing the number of categories corresponds to taking into account human interpretations more substantially during evaluation. One can further argue that true interpretability scores of the embeddings (i.e. scores from an all-comprehensive dataset) should be even larger than those presented in Table III. However, it can also be noticed that the increase in the interpretability scores

TABLE III
AVERAGE INTERPRETABILITY SCORES (%) FOR $\lambda = 5$

Category Coverage (%)		Number of Categories: 30	50	70	90	110
40	Random	4.9	5.5	6.0	6.4	6.7
	GloVe	5.6	6.8	7.7	8.3	8.9
	$\mathcal{I}^*$	25.9	33.6	40.2	44.8	49.1
	$\mathcal{I}$	34.2	45.2	55.5	62.9	69.2
60	Random	4.5	4.9	5.3	5.6	5.8
	GloVe	6.7	7.8	9.0	9.7	10.2
	$\mathcal{I}^*$	27.6	35.8	42.4	47.7	51.6
	$\mathcal{I}$	36.1	48.4	59.0	67.0	72.8
80	Random	4.2	4.6	4.9	5.1	5.3
	GloVe	7.6	8.9	9.7	10.4	11.0
	$\mathcal{I}^*$	30.2	31.1	43.2	48.1	52.0
	$\mathcal{I}$	39.8	50.7	60.1	67.4	73.2
100	Random	4.3	4.6	4.8	5.0	5.1
	GloVe	8.4	9.8	10.8	11.4	12.0
	$\mathcal{I}^*$	30.3	37.7	43.4	48.1	51.5
	$\mathcal{I}$	38.9	49.9	59.0	65.7	71.3

Results are averaged across 10 independent selections of categories for each category coverage.

of the GloVe and random embedding spaces gets smaller for larger number of categories. Thus, there is diminishing returns to increasing number of categories in terms of interpretability. Another important observation is that the interpretability scores of $\mathcal{I}$ and $\mathcal{I}^*$ are more sensitive to number of categories in the dataset than the GloVe or random embeddings. This can be attributed to the fact that $\mathcal{I}$ and $\mathcal{I}^*$ comprise dimensions that correspond to SEMCAT categories, and that inclusion or exclusion of these categories more directly affects interpretability.

In contrast to the category coverage, the effects of within-category word coverage on interpretability scores can be more complex. Starting with few words within each category, increasing the number of words is expected to more uniformly sample from the word distribution, more accurately reflect the semantic relations within each category and thereby enhance interpretability scores. However, having categories over-abundant in words might inevitably weaken semantic correlations among them, reducing the discriminability of the categories and interpretability of the embedding. Table III shows that, interestingly, changing the category coverage has different effects on the interpretability scores of different types of embeddings. As category word coverage increases, interpretability scores for random embedding gradually decrease while they monotonically increase for the GloVe embedding. For semantic spaces $\mathcal{I}$ and $\mathcal{I}^*$, interpretability scores increase as the category coverage increases up to 80% of that of SEMCAT, then the scores decrease. This may be a result of having too comprehensive categories as argued earlier, implying that categories with coverage of around 80% of SEMCAT are better suited for measuring interpretability. However, it should be noted that the change in the interpretability scores for different word coverages might be effected by non-ideal subsampling of category words. Although our word sampling method, based on words' distances to category centers, is expected to generate categories that are represented better compared to random sampling of category words, category representations might be suboptimal compared to human designed categories.

V. DISCUSSION AND CONCLUSION

In this paper, we propose a statistical method to uncover the latent semantic structure in dense word embeddings. Based on a new dataset (SEMCAT) we introduce that contains more than 6,500 words semantically grouped under 110 categories, we provide a semantic decomposition of the word embedding dimensions and verify our findings using qualitative and quantitative tests. We also introduce a method to quantify the interpretability of word embeddings based on SEMCAT that can replace the word intrusion test that relies heavily on human effort while keeping the basis of the interpretations as human judgements.

Our proposed method to investigate the hidden semantic structure in the embedding space is based on calculation of category weights using a Bhattacharya distance metric. This metric implicitly assumes that the distribution of words within each embedding dimension is normal. Our statistical assessments indicate that the GloVe embedding space considered here closely follows this assumption. In applications where the embedding method yields distributions that significantly deviate from a normal distribution, nonparametric distribution metrics such as Spearman's correlation could be leveraged as an alternative. The resulting category weights can seamlessly be input to the remaining components of our framework.

Since our proposed framework for measuring interpretability depends solely on the selection of the category words dataset, it can be used to directly compare different word embedding methods (e.g., GloVe, word2vec, fasttext) in terms of the interpretability of the resulting embedding spaces. A straightforward way to do this is to compare the category weights calculated for embedding dimensions across various embedding spaces. Note, however, that the Bhattacharya distance metric for measuring the category weights does not follow a linear scale and is unbounded. For instance, consider a pair of embeddings with category weights 10 and 30 versus another pair with weights 30 and 50. For both pairs, the latter embedding can be deemed more interpretable than the former. Yet, due to the gross nonlinearity of the distance metric, it is challenging to infer whether a 20-unit improvement in the category weights corresponds to similar levels of improvement in interpretability across the two pairs. To alleviate these issues, here we propose an improved method that assigns normalized interpretability scores with an upper bound of 100%. This method facilitates interpretability assessments and comparisons among separate embedding spaces.

The results reported in this study for semantic analysis and interpretability assessment of embeddings are based on SEMCAT. SEMCAT contains 110 different semantic categories where average number of words per category is 91 rendering SEMCAT categories quite comprehensive. Although the HyperLex dataset contains a relatively larger number of categories (1399), the average number of words per category is only 2, insufficient to accurately represent semantic categories. Furthermore, while HyperLex categories are constructed based on a single type of relation among words (hyperonym-hyponym), SEMCAT is significantly more comprehensive since many categories include words that are grouped based on diverse types of relationships

that go beyond hypernym-hyponym relations. Meanwhile, the relatively smaller number of categories in SEMCAT is not considered a strong limitation, as our analyses indicate that the interpretability levels exhibit diminishing returns when the number of categories in the dataset are increased and SEMCAT is readily yielding near optimal performance. That said, extended datasets with improved coverage and expert labeling by multiple observers would further improve the reliability of the proposed approach. To do this, a synergistic merge with existing lexical databases such as WordNet might prove useful.

Methods for learning dense word embeddings remain an active area of NLP research. The framework proposed in this study enables quantitative assessments on the intrinsic semantic structure and interpretability of word embeddings. Providing performance improvements in other common NLP tasks might be a future study. Therefore, the proposed framework can be a valuable tool in guiding future research on obtaining interpretable yet effective embedding spaces for many NLP tasks that critically rely on semantic information. For instance, performance evaluation of more interpretable word embeddings on higher level NLP tasks (i.e. sentiment analysis, named entity recognition, question answering) and the relation between interpretability and NLP performance can be worthwhile.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their constructive and helpful comments that have significantly improved this paper.

REFERENCES

[1] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

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

[3] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Proc. Artif. Intell. Statist.*, 2012, pp. 127–135.

[4] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.

[5] J. R. Firth, "A synopsis of linguistic theory, 1930–1955," in *Stud. Linguistic Anal.* Oxford, U.K.: Blackwell, 1957.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, p. 391, 1990.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.

[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[10] C.-C. Lin, W. Ammar, C. Dyer, and L. Levin, "Unsupervised POS induction with word embeddings," in *Proc. 2015 Conf. North Amer. Ch. Assoc. Comput. Linguistics: Human Lang. Technol.*, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 1311–1316.

[11] S. K. Sienčnik, "Adapting word2vec to named entity recognition," in *Proc. Nordic Conf. Comput. Linguistics*, Linköping University Electronic Press, Vilnius, Lithuania, May 11–13, 2015, vol. 109, pp. 239–243.

[12] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Embeddings for word sense disambiguation: An evaluation study," in *Proc. Assoc. Comput. Linguistics*, 2016, vol. 1, pp. 897–907.

[13] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," in *Proc. Empirical Methods Natural Lang. Process.*, 2017, pp. 534–539.

[14] L. K. Şenel, V. Yücesoy, A. Koç, and T. Çukur, "Measuring cross-lingual semantic similarity across european languages," in *Proc. 40th Int. Conf. Telecommun. Signal Process.*, 2017, pp. 359–363.

[15] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proc. Assoc. Comput. Linguistics*, 2014, pp. 302–308.

[16] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behav. Res. Methods, Instrum. Comput.*, vol. 28, no. 2, pp. 203–208, 1996.

[17] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," preprint arXiv:1606.08813, 2016.

[18] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Intell. Res.*, vol. 49, no. 2014, pp. 1–47, 2014.

[19] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Comput. Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[20] G. Murphy, *The Big Book of Concepts*. Cambridge, MA, USA: MIT Press, 2004.

[21] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. Int. Conf. Neural Inf. Process. Systems*, 2009, pp. 288–296.

[22] B. Murphy, P. Talukdar, and T. Mitchell, "Learning effective and interpretable semantic models using non-negative sparse embedding," in *Proc. Comput. Linguistics*, 2012, pp. 1933–1950.

[23] H. Luo, Z. Liu, H.-B. Luan, and M. Sun, "Online learning of interpretable word embeddings," in *Proc. Empirical Methods Natural Lang. Process.*, 2015, pp. 1687–1692.

[24] A. Fyshe, P. P. Talukdar, B. Murphy, and T. M. Mitchell, "Interpretable semantic vectors from a joint model of brain-and text-based meaning," in *Proc. Assoc. Comput. Linguistics*, NIH Public Access, 2014, vol. 2014, pp. 489–499.

[25] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "Linear algebraic structure of word senses, with applications to polysemy," preprint arXiv:1601.03764, 2016.

[26] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. Smith, "Sparse overcomplete word vector representations," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics and 7th Int. Joint Conf. Natural Lang. Process.*, Association for Computational Linguistics, Beijing, China, vol. 1, Jul. 2015, pp. 1491–1500.

[27] A. Zobnin, "Rotations and interpretability of word embeddings: The case of the Russian language," preprint arXiv:1707.04662, 2017.

[28] S. Park, J. Bak, and A. Oh, "Rotated word vector representations and their interpretability," in *Proc. Empirical Methods Natural Lang. Process.*, 2017, pp. 401–411.

[29] K.-R. Jang and S.-H. Myaeng, "Elucidating conceptual properties from word embeddings," in *Proc. Sense, Concept, Entity Represent. Appl.*, 2017, pp. 91–95.

[30] I. Vulić, D. Gerz, D. Kiela, F. Hill, and A. Korhonen, "Hyperlex: A large-scale evaluation of graded lexical entailment," *Comput. Linguistics*, vol. 43, no. 4, pp. 781–835, 2017.

[31] A. Gladkova, A. Drozd, and C. Center, "Intrinsic evaluations of word embeddings: What can we do better?" in *Proc. 1st Workshop Eval. Vector Space Represent. NLP*, 2016, pp. 36–42.

[32] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distribution," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

Authors' photographs and biographies not available at the time of publication.