Interpreting Word Embeddings with Eigenvector Analysis

Jamin Shin **Andrea Madotto** **Pascale Fung**
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
{jay.shin, amadotto}@connect.ust.hk
pascale@ece.ust.hk

Abstract

Dense word vectors have proven their values in many downstream NLP tasks over the past few years. However, the dimensions of such embeddings are not easily interpretable. Out of the d-dimensions in a word vector, we would not be able to understand what high or low values mean. Previous approaches addressing this issue have mainly focused on either training sparse/non-negative constrained word embeddings, or post-processing standard pre-trained word embeddings. On the other hand, we analyze conventional word embeddings trained with Singular Value Decomposition, and reveal similar interpretability. We use a novel eigenvector analysis method inspired from Random Matrix Theory and show that semantically coherent groups not only form in the row space, but also the column space. This allows us to view individual word vector dimensions as human-interpretable semantic features.

1 Introduction

Understanding words has a fundamental impact on many natural language processing tasks, and has been modeled with the Distributional Hypothesis [1]. Dense d-dimensional vector representations of words created from this model are often referred to as *word embeddings*, and have successfully captured similarities between words, such as *word2vec* and *GloVe* [2, 3]. They have also been applied to downstream NLP tasks as word representation features, ranging from sentiment analysis to machine translation [4, 5].

Despite their widespread popularity in usage, the dimensions of these word vectors are difficult to interpret [6]. Consider $\mathbf{w}_{president} = [0.1, 2.4, 0.3]$ as the 3-dimensional vector of "president" from *word2vec*. In this 3-dimensional space (or the row space), semantically similar words like "minister" and "president" are closely located. However, it is unclear what the dimensions represent, as *we do not know the meaning of the 2.4 in* $\mathbf{w}_{president}$. It is difficult to answer questions like '*what is the meaning of high and low values in the columns of* $\mathbf{W}$' and '*how can we interpret the dimensions of word vectors*'. To address this problem, previous literature focused on the column space by either training word embeddings with sparse and non-negative constraints [7–9], or post-processing pre-trained word embeddings [6, 10, 11]. We instead investigate this problem from a random matrix perspective.

In our work, we analyze the eigenvectors of word embeddings obtained with *truncated Singular Value Decomposition (SVD)* [12, 13] of the *Positive Pointwise Mutual Information (PPMI)* matrix [14]. Moreover, we compare this analysis with the row and column space analysis of *Skip Gram Negative Sampling (SGNS)*, a model used to train *word2vec* [15]. From the works of [16] proving that both SVD and SGNS factorizes and approximates the same matrix, we hypothesize that a study of the principal eigenvectors of the PPMI matrix reflects the information contained in SGNS.

32nd Conference on Neural Information Processing Systems (NIPS 2018), IRASL workshop Montréal, Canada.

Contributions: Without requiring any constraints or post-processing, we show that the dimensions of word vectors can be interpreted as semantic features. In doing so, we also introduce novel word embedding analysis methods inspired by the literature of eigenvector analysis techniques from Random Matrix Theory.

2 Related Work

Recently, there have been several works that have shown similar results in semantic grouping among the column values. Several of these algorithms proposed to train non-negative sparse interpretable word vectors [7–9, 17].

Furthermore, [6] also proposed methods to post-process pre-trained word vectors with non-negativity and sparsity constraints. However, their vectors were optionally binarized, which is difficult to interpret intensity than real-values. [10] has proposed to overcome these limitations by simply training a rotation matrix to transform pre-trained *word2vec* and *GloVe*, without being sparse or binary. Finally, [11] post-trained the pre-trained word embeddings with k-sparse autoencoders with similar constraints to [6].

While these methods were able to successfully achieve interpretability in the column space evaluated with word intrusion detection tests, they either enforced sparsity and non-negativity constraints, or required extensive post-processing. Furthermore, they focused less on the analysis and discussion on the actual meanings of the columns despite their pursuit of interpretable dimensions. Hence, in our work, we put more emphasis on such implications with conventional algorithms without any extra constraints or post-processing steps.

3 Methodology

3.1 Notations

We define the Positive Pointwise Mutual Information (PPMI) matrix as $\mathbf{M}^{\text{PPMI}}$, the set of unique words as vocabulary V, and word embedding matrices created from SVD and SGNS as $\mathbf{W}^{\text{SVD}}$ and $\mathbf{W}^{\text{SGNS}}$. The k-th largest eigenvalue and corresponding eigenvector of $\mathbf{M}^{\text{PPMI}}$ are denoted as λ^k and $\mathbf{u}^k \in \mathbb{R}^{|V|}$, and the k-th column of $\mathbf{W}^{\text{SGNS}}$ as $\mathbf{v}^k \in \mathbb{R}^{|V|}$. The word vectors are denoted $\mathbf{w}^{\text{SVD}}_{word}$ or $\mathbf{w}^{\text{SVD}}_{word}$, but when context is clear or does not matter, we simply use $\mathbf{w}_{word}$. Note that we often use the term "eigen" when and "singular" interchangeably because $\mathbf{M}^{\text{PPMI}}$ is defined as a square matrix.

3.2 Positive Pointwise Mutual Information (PPMI) Matrix

Each entry of a co-occurrence matrix $\mathbf{M}$ represents the co-occurrence counts of words w_i and c_j in all documents in the corpus. However, raw co-occurrence counts have been known to underperform than other transformed variants [16]. Pointwise Mutual Information (PMI) [14] instead transforms matrix by measuring the log ratio between the joint probability of w and c when assuming independence of the two and not.

$$PMI(w,c) = \log \frac{\hat{P}(w,c)}{\hat{P}(w)\hat{P}(c)} = \log \frac{\#(w,c)|D|}{\#(w)\#(c)}$$

The problem of this association measure is when dealing with never observed pairs which result in $PMI(w,c) = \log 0$. To cope with such, Positive Pointwise Mutual Information has been used to map all negative values to 0 from the intuition that positive associations are often more informative in downstream NLP tasks [16].

$$PPMI(w,c) = \max(PMI(w,c), 0)$$

3.3 Truncated Singular Value Decomposition (SVD)

Truncated SVD (we will further refer this as simply SVD), which is equivalent to maximum variance Principal Component Analysis (PCA) and has been popularized by Latent Semantic Analysis (LSA) [13], factorizes the PPMI matrix as $\mathbf{M}^{\text{PPMI}} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^{\text{T}}$ and truncates to d dimensions. Following the works of [18], the word embedding matrix is taken as $\mathbf{W} = \mathbf{U}_d$, instead of the more “standard" eigenvalue weighting $\mathbf{W} = \mathbf{U}_d \cdot \mathbf{S}$. We discuss the effect of this in Section 6.2.

3.4 Skip-Gram with Negative Sampling (SGNS)

Unlike PPMI and SVD which gives exact solutions, the *word2vec* Skip-Gram model, proposed by [2], trains two randomly initialized word embedding matrices $\mathbf{W}$ and $\mathbf{C}$ with a neural network.

$$P(\mathbf{C}_j|\mathbf{W}_i) = Softmax(\mathbf{W}_i \cdot \mathbf{C}_j), \qquad \text{where } Softmax(\mathbf{W}_i \cdot \mathbf{C}_j) = \frac{e^{\mathbf{W}_i \cdot \mathbf{C}_j}}{\sum_k e^{\mathbf{W}_i \cdot \mathbf{C}_k}}$$

The intuition is to basically maximize the dot product between "similar" word and context pairs, and minimize the dot product between wrong pairs. The *Softmax* function is simply a generalized version of the logistic function to multi-class scenario. However, the normalization constant which computes the exponentials of all context words, is very computationally expensive when the vocabulary size is large. Hence, [15] proposed Skip Gram with Negative Sampling (SGNS) to simplify the objective using negative sampling.

3.5 Eigenvector Analysis Methods

We borrow intuitions from the Random Matrix Theory literature to analyze eigenvectors of $\mathbf{M}^{\text{PPMI}}$. We analyze the distributions of eigenvectors, calculate the Inverse Participation Ratios (IPR) to quantify the ratio of significant elements and measure structural sparsity, and qualitatively interpret the significant elements.

Distribution of Eigenvector: The empirical distribution of eigenvector elements $\mathbf{u}^k$ is compared with a Normal distribution $N(\mu_{\mathbf{u}^k}, \sigma^2_{\mathbf{u}^k})$ to measure normality of the eigenvectors, where $\mu_{\mathbf{u}^k}, \sigma^2_{\mathbf{u}^k}$ refer to the mean and variance of $\mathbf{u}^k$. [19] have shown that eigenvectors deviating from Gaussian contain genuine correlation between stocks, while also revealing a global bias that represented newsbreaks influencing all stocks. We search for similar patterns in Section 5.1.

Inverse Participation Ratio: The Inverse Participation Ratio (IPR) of $\mathbf{u}^k$, denoted as I^k, quantifies the inverse ratio of significant elements in the eigenvector $\mathbf{u}^k$ [19–21].

$$I^k \triangleq \sum_{i=1}^{|V|} [\mathbf{u}_i^k]^4,$$

where $\mathbf{u}_i^k$ is the i-th element of $\mathbf{u}^k$. The intuition of IPR can be illustrated with two extreme cases. First, if all elements of $\mathbf{u}^k$ have same values $1/\sqrt{|V|}$, then I^k is simply $1/|V|$, with reciprocal $1/I^k$ being $|V|$. This means that all $|V|$ elements contribute similarly. On the other hand, a one-hot vector with only one element as one, and the rest as zero, $\mathbf{u}^k$ will have an IPR value of one (also same for reciprocal). Hence, the reciprocal, $1/I^k$, measures the ratio of significant participants in $\mathbf{u}^k$. In short, the larger the I^k, the smaller the ratio of participation, and the sparser the vector, in turn, reflecting structural sparsity of $\mathbf{u}^k$. Furthermore, as $1/I^k \in [1, |V|]$, dividing this reciprocal with $|V|$ will yield the sparsity of a given vector $\mathbf{u}^k \in \mathbb{R}^{|V|}$.

Visualization of Top Eigenvector elements: As $\mathbf{u}^k, \mathbf{v}^k \in \mathbb{R}^{|V|}$, we can map each index of the vectors to a *word* in the vocabulary V. Hence, we investigate the dimensions and their indices (or *words*) with the largest absolute values and search for semantic coherence. Similar approaches with financial data have shown to group stocks from same industries or nearby regions [19], and with genetic data, revealed important co-evolving genes in gene co-expression networks [20].

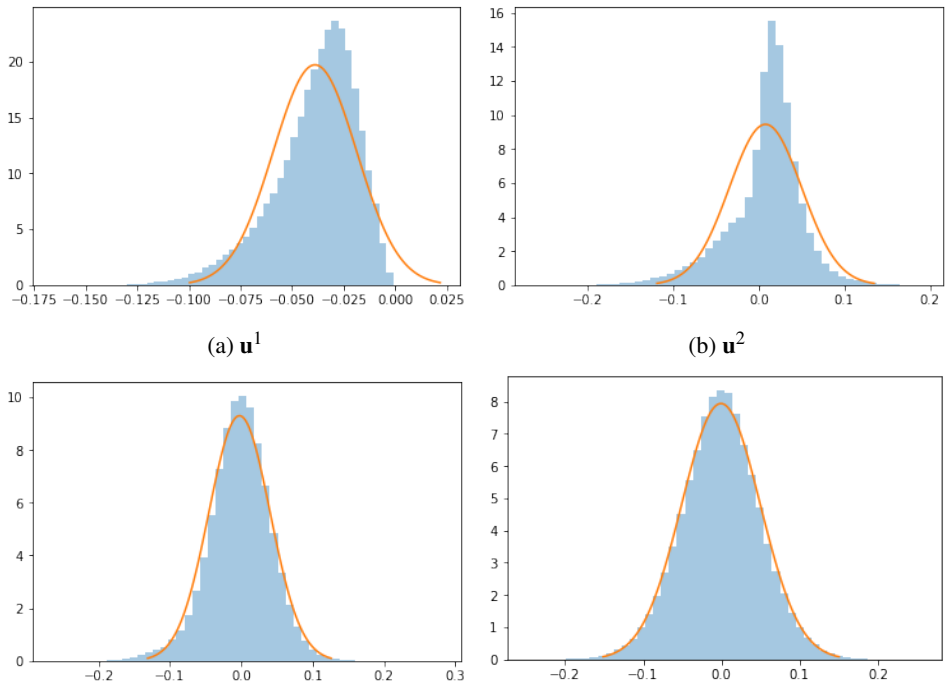
4 Experimental Setup

4.1 Training

English Wikipedia We use the English Wikipedia dump[1] cleaned by adapting Matt Mahoney's Perl script[2], which has also been used by [2]. Removing most of the noisy non-alphanumerics, such as XML tags, the dataset size effectively reduced from approximately 66GB to 25GB, containing around 3.4B tokens. The vocabulary size is approximately 346K as we only consider words with at least 100 occurrences.

[1] `https://dumps.wikimedia.org/enwiki/20180420/`

[2] At the bottom of `http://mattmahoney.net/dc/textdata.html`



(a) $\mathbf{u}^1$ (b) $\mathbf{u}^2$

(c) $\mathbf{u}^{100}$ (d) $\mathbf{u}^{500}$

Figure 1: Eigenvector distributions of $\mathbf{u}^1$, $\mathbf{u}^2$, $\mathbf{u}^{100}$, $\mathbf{u}^{500}$ ($\mathbf{u}^1$ is the largest eigenvector). Solid curves are Gaussian.

SGNS and SVD We adapt the code from the *hyperwords*[3] released by [18] to train both $\mathbf{W}^{\text{SVD}}$ and $\mathbf{W}^{\text{SGNS}}$. Our code is publicly available online[4]. For $\mathbf{W}^{\text{SGNS}}$, we set negative sampling as 5. For both, we set a context window size of 2 (taking 5 words as context) and embedding dimension $d = 500$.

5 Results

5.1 Distribution of Eigenvector Elements

From Figure 1, we can see that eigenvectors corresponding to the larger eigenvalues such as $\mathbf{u}^1$ or $\mathbf{u}^2$ clearly deviate from a Gaussian distribution, and so do $\mathbf{u}^{100}$ and $\mathbf{u}^{500}$, but less. This shows us that the eigenvectors are not random and contain meaningful correlations. It is expected to see such pattern because these vectors are the principal eigenvectors.

On a more interesting note, $\mathbf{u}^1$ not only significantly deviates from a normal distribution, but also has only non-zero negative values as its elements, and no other eigenvectors have shown this behavior. This suggests that this particular eigenvector could represent a common bias that affects all "words", as it captured the effect of news outbreaks for stock prices in [19]. We revisit the interpretation of this observation in Section 6.1.

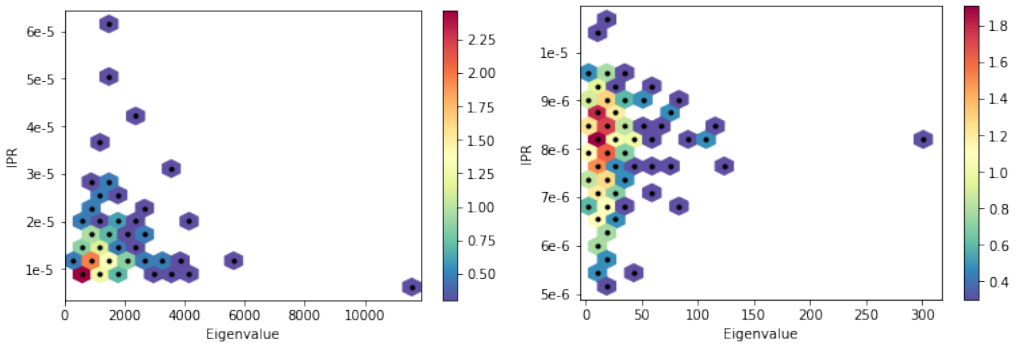
5.2 Inverse Participation Ratio

Figure 2 illustrates the IPR of $\mathbf{u}^k$ plotted against the corresponding eigenvalue λ^k, and vice versa for $\mathbf{v}^k$. From the plot, we can clearly see that the eigenvectors of $\mathbf{W}^{\text{SVD}}$ have approximately 10x higher IPR values than those of $\mathbf{W}^{\text{SGNS}}$, meaning that the vectors are much sparser for $\mathbf{W}^{\text{SVD}}$.

From Figure 2a, we can see that the largest eigenvector has the smallest IPR of 0.000006, and the reciprocal $1/I^k$ divided by $|V|$, yields 48%, while the same for the largest I^k gave around 4.7%. The

[3] `http://bitbucket.org/omerlevy/hyperwords`

[4] `https://github.com/HLTCHKUST/eigenvector-analysis`



(a) $\mathbf{W}^{\text{SVD}}$ (b) $\mathbf{W}^{\text{SGNS}}$

Figure 2: Inverse participation ratios. The more red the dots are, more points are concentrated.

$\mathbf{u}^1$	$\mathbf{u}^4$	$\mathbf{u}^7$	$\mathbf{u}^8$	$\mathbf{u}^{14}$	$\mathbf{u}^{121}$
lastly	molly	determinants	shyam	famille	jays
outset	sally	biochemical	sanjeev	vrier	strikeouts
ostensibly	toby	intrinsic	meera	autour	halladay
curiously	maggie	qualitative	anupama	naissance	hitters
actuality	valentine	elucidated	deepa	rique	buehrle
crucially	jenny	analytical	rajkumar	diteur	batters
theirs	tracy	psychological	manju	octobre	pitching
importantly	lucy	unger	uday	chambre	phillies
thankfully	carrie	ehrlich	chitra	lettre	rbis
regrettably	elliot	quantitative	vinod	campagne	astros
ironically	susie	integrative	archana	jeune	diamondbacks
aforementioned	laurie	extrinsic	bhanu	jours	homers
paradoxically	cooper	nagel	santosh	septembre	hitless
oftentimes	jill	methodologies	rajesh	enfance	orioles
doubtless	kitty	exogenous	ashok	plon	podsednik
unsurprisingly	charlie	underneath	munna	affaire	baserunners
connelly	shirley	translational	suman	cembre	hitter
merrick	hannah	kuhn	komal	royaume	sox
invariably	annie	functional	subhash	propos	pettitte
dunning	elaine	schweitzer	usha	juin	vizquel
Transition	First Names	Science	Indian Names	French	Baseball

Table 1: Top participants of eigenvectors (dimensions with highest magnitudes) of $\mathbf{W}^{\text{SVD}}$ form semantically coherent groups. $\mathbf{u}^{14}$ and $\mathbf{u}^{121}$ are eigenvectors with large IPR value, while the remaining are corresponding eigenvectors of the largest eigenvalues.

mean value of $1/I^k$ divided by $|V|$, across all eigenvectors was 27.5% indicating that there exists some sparse structure within the eigenvectors of $\mathbf{W}^{\text{SVD}}$. On the other hand, Figure 2b shows that mean for $\mathbf{v}^k$ was around 36%, meaning that column vectors of $\mathbf{W}^{\text{SGNS}}$ are generally denser and less structured. Such discrepancy in structural sparsity motivates us to analyze the eigenvectors of $\mathbf{W}^{\text{SVD}}$ in depth.

6 Analysis and Discussion

6.1 Column Space Analysis

Based on the results of previous sections, we further examine the top elements of the eigenvectors by sorting their absolute values in decreasing order. Table 1 shows interesting results as the significant dimensions or their corresponding “words" of each eigenvector, in general, form semantically or syntactically coherent groups. For instance, $\mathbf{u}^{14}$ groups French words together and $\mathbf{u}^{121}$ shows

$\mathbf{u}^{42}$	$\mathbf{u}^{50}$	$\mathbf{u}^{14}$	$\mathbf{u}^{101}$	$\mathbf{u}^{75}$	$\mathbf{u}^{121}$
stani	bandeira	famille	seon	bucharest	jays
kne	concei	vrier	hyeon	lcescu	strikeouts
vukovi	nio	autour	seung	ional	halladay
kovi	jardim	naissance	seong	ntul	hitters
vuk	velho	rique	hwang	napoca	buehrle
kova	visconde	diteur	choi	editura	batters
popovi	pessoa	octobre	sik	iancu	pitching
inovi	domingos	chambre	kyung	romanian	phillies
uro	pinheiro	lettre	kwang	mihai	rbis
mileti	branco	campagne	jeong	institutul	astros
novakovi	ncio	jeune	yeong	tilor	diamondbacks
anovi	trindade	jours	taek	traian	homers
filipovi	carmo	septembre	gyeong	ianu	hitless
petrovi	cio	enfance	seok	pentru	orioles
radi	ssimo	plon	choe	tefan	podsednik
evi	penha	affaire	ryong	muzeul	baserunners
knez	paulo	cembre	hee	gheorghe	hitter
martinovi	cavalo	royaume	gwang	bucuresti	sox
vuka	marinho	propos	cheol	biserica	pettitte
veljko	neves	juin	myeong	craiova	vizquel
Slavic	Brazillian	French	Korean Names	Romanian	Baseball

Table 2: Top participants of eigenvectors (dimensions with highest magnitudes) of $\mathbf{W}^{\text{SVD}}$ with largest IPR values (highest sparsity) in decreasing order.

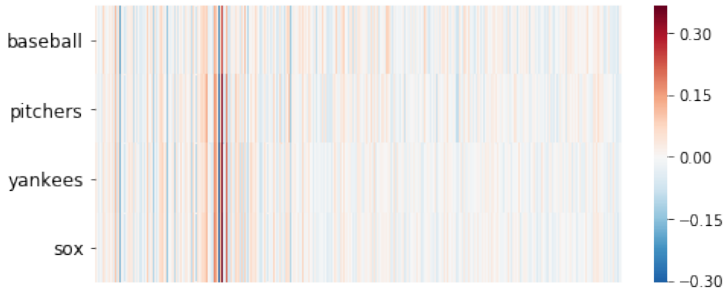
$\mathbf{u}^{53}$	$\mathbf{u}^{337}$	$\mathbf{v}^{447}$	$\mathbf{v}^{229}$
located	vueling	garabed	itch
near	tuifly	hagopian	negros
connecting	eurowings	activist	sulu
situated	tunisair	abrahamyan	sulawesi
connects	fiumicino	marash	fukuyama
bandeira	interjet	voices	ozu
trindade	transavia	papazian	dimetrodon
penha	wizz	vardapet	occidental
concei	easyjet	erden	paralysis
velho	volotea	documentaries	paths
$\mathbf{W}^{\text{SVD}}$	$\mathbf{W}^{\text{SVD}}$	$\mathbf{W}^{\text{SGNS}}$	$\mathbf{W}^{\text{SGNS}}$

Table 3: Top participants of the salient columns of the word vector for "airport."

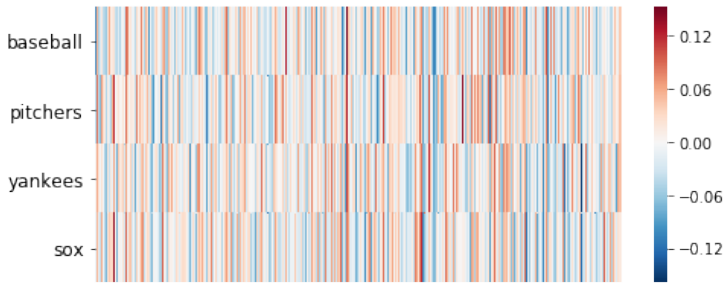
baseball related words. Some words from $\mathbf{u}^{121}$ initially seem irrelevant to baseball. However, "buehrle" is a baseball player, "rbis" stand for "Run Batted Ins", and "astros" is a baseball team name from Houston. Meanwhile, the words grouped in $\mathbf{u}^{1}$, the largest eigenvector, could explain the bias we mentioned in Section 5.1. The significant participants tend to be *strong* transition words that are used often for dramatic effects, such as "importantly" or "crucially". Evidently, these words increase the intensity of the context.

Moreover, while it was originally hypothesized that the largest principal eigenvectors would capture some semantic relationship, the 121th vector $\mathbf{u}^{121}$ show surprisingly focused and narrow semantic grouping related to baseball. Further investigation reveals that $\mathbf{u}^{121}$ has one of the highest IPR values, hence being one of the most sparse vectors. We verify similar trends in other eigenvectors with high IPR values as shown in Table 2. An interesting pattern arises here, in which the sparser eigenvectors tend to capture more distinct and rare features such as foreign names or languages, or topics like baseball.

Furthermore, we compare the column space analysis on $\mathbf{W}^{\text{SVD}}$ and $\mathbf{W}^{\text{SGNS}}$. Consider the word vector $\mathbf{w}_{airport}$ for the word "airport." We choose the *salient* dimensions, which are the largest elements, of $\mathbf{w}_{airport}$, and investigate the significant elements of those chosen dimensions (columns). Table



(a) Word vectors from $\mathbf{W}^{\text{SVD}}$



(b) Word vectors from $\mathbf{W}^{\text{SGNS}}$

Figure 3: Comparison of the representations of four baseball related words in $\mathbf{W}^{\text{SVD}}$ and $\mathbf{W}^{\text{SGNS}}$, where the rows are embedding dimensions. It is clear that $\mathbf{W}^{\text{SVD}}$ shows similar representations.

3 shows that the columns from $\mathbf{W}^{\text{SVD}}$ display semantic coherence while those from $\mathbf{W}^{\text{SGNS}}$ seem random. $\mathbf{u}^{53}$ groups words that are related to the location of the airports. For example, one could say "Trindade station connects with the airport." Similarly, $\mathbf{u}^{337}$ groups famous airline companies together, while "fiumicino" is a famous airport in Italy.

6.2 Word Embedding Dimensions as Interpretable Features

Sections 5.1 and 5.2 revealed that the eigenvectors contain genuine correlation and structure in the column space. We further show in Section 6.1 that semantically coherent words form groups of significant participants in each eigenvector. Now we can answer the questions we asked earlier.

***What is the meaning of high and low values in the columns of* W?** If word vector $\mathbf{w}$ from $\mathbf{W}^{\text{SVD}}$ has a high absolute value in column k, it means that the word is relevant to the semantic group formed in $\mathbf{u}^k$. For example, the words from Figure 3a have highest values in column $k = 121$, in which $\mathbf{u}^{121}$ represents a semantic group related to baseball, as shown in Table 1.

***How can we interpret the dimensions of word vectors*?** The answer to this question follows naturally. As the salient dimensions represent relevant semantic groups, we can view the dimensions of $\mathbf{w}$ as semantic features. This view is in line with the Topic Modeling literature, in which words and documents are clustered into distinct latent topics. Hence, we can also see the word embedding dimensions as latent topics that can be interpretable.

It can be easily seen from Figure 3b that similar words do not show any interpretable similarity in their $\mathbf{W}^{\text{SGNS}}$ representations, despite being nearest neighbors in the row space. On the other hand, it is very clear from Figure 3a that similar words have similar representations, or feature vectors. We thus empirically verify that the dimensions of the row vectors can be viewed as semantic or syntactic features. Finally, the structural sparsity discovered with the IPR is further confirmed by contrasting Figures 3a and 3b. It is clearly visible that the the vectors from SVD are much sparser than from SGNS.

Effect of Eigenvalue Weighting: As mentioned in Section 3, weighting with the eigenvalues essentially scales each feature column by the corresponding eigenvalues. Such process can be viewed

as simply incorporating a prior, and does not hurt the interpretability. However, as [18] showed that eigenvalue weighting decreases the performance of downstream NLP tasks, we can assume that either the prior is wrong, or too strong. In fact, in many cases, the largest eigenvalues are often order of magnitude larger than others, which can explain why not weighting the word embeddings with their corresponding eigenvalues would work better.

7 Conclusion

In this work, we analyzed the eigenvectors, or the column space, of the word embeddings obtained from the Singular Value Decomposition of PPMI matrix. We revealed that the significant participants of the eigenvectors form semantically coherent groups, allowing us to view each word vector as an interpretable feature vector composed of semantic groups. These results can be very useful in error analysis in downstream NLP tasks, or cherry-picking useful feature dimensions to easily create compressed and efficient task-specific embeddings. Future work will proceed in this direction on applying interpretability to practical usage.

Acknowledgments

This work is partially funded by ITS/319/16FP of Innovation Technology Commission, HKUST 16248016 of Hong Kong Research Grants Council.

References

[1] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1162.

[4] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1181.

[5] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *Sixth International Conference on Learning Representations (ICLR 2018)*, 2018.

[6] Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-1144.

[7] Brian Murphy, Partha Talukdar, and Tom Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING 2012*, pages 1933–1950, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL http://www.aclweb.org/anthology/C12-1118.

[8] Hongyin Luo, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Online learning of interpretable word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1687–1692, 2015.

[9] Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. Sparse word embeddings using l1 regularized online learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2915–2921. AAAI Press, 2016.

[10] Sungjoon Park, JinYeong Bak, and Alice Oh. Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 401–411, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1041.

[11] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. Spine: Sparse interpretable neural embeddings. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[12] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[13] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[14] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March 1990. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=89086.89095.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Third International Conference on Learning Representations (ICLR 2013)*, 2013.

[16] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[17] Thomas Kober, Julie Weeds, Jeremy Reffin, and David Weir. Improving sparse word representations with distributional inference for semantic composition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1702, 2016.

[18] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. ISSN 2307-387X. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570.

[19] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A Nunes Amaral, Thomas Guhr, and H Eugene Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126, 2002.

[20] Sarika Jalan, Norbert Solymosi, Gábor Vattay, and Baowen Li. Random matrix analysis of localization properties of gene coexpression network. *Physical Review E*, 81(4):046118, 2010.

[21] Priodyuti Pradhan, Alok Yadav, Sanjiv K Dwivedi, and Sarika Jalan. Optimized evolution of networks for principal eigenvector localization. *Physical Review E*, 96(2):022312, 2017.