AlQalam
Alq J Med App Sci
https://journal.utripoli.edu.ly/index.php/Alqalam/index eISSN 2707-7179

Original article

Analyzing Semantic Properties of Word Embeddings Using Eigenvectors

Bassma Abdlrazg*, Hanan Atetalla

Department Mathematics, University of Omar AL-Mokhtar, ALbaida, Libya.

ARTICLE INFO

Corresponding Email. *bassma.abraheem@omu.edu.ly*

Received: *14-08-2024*
Accepted: *21-10-2024*
Published: *28-10-2024*

Keywords. *Word Vectors, Word Embeddings, Word2Vec, Embedding Space, Eigenvectors.*

ABSTRACT

Dense word vectors have demonstrated their efficacy in several downstream natural language processing (NLP) tasks in recent years. Nevertheless, the interpretability of these embeddings' dimensions remains challenging. In this paper, we investigate how eigenvectors can reveal different semantic properties captured by word embedding models. Hence, we train word embeddings (e.g., Word2Vec) on English Wikipedia corpus to analyze the top eigenvectors, identify specific semantic properties (e.g., sentiment, formality) associated with each, and explore how these properties are encoded in the embedding space. This paper also discussed the limitations and potential benefits of this approach compared to other methods for analyzing word embeddings.

Cite this article. *Abdlrazg B, Atetalla H. Analyzing Semantic Properties of Word Embeddings Using Eigenvectors. Alq J Med App Sci. 2024;7(4):1088-1093. https://doi.org/10.54361/ajmas.247424*

INTRODUCTION

Comprehending words plays a crucial role in various natural language processing activities, and has been conceptualized through the Distributional Hypothesis. Word embeddings, which are dense d-dimensional vector representations of words derived from this concept, are commonly known for capturing semantic similarities between words, as demonstrated by word2vec and GloVe [1-6]. These embeddings have proven to be useful in a wide range of NLP tasks, serving as word representation features in applications such as sentiment analysis and machine translation.

In spite of their widespread usage, the interpretation of the dimensions of word vectors remains challenging [6]. Let's consider the vector $w_{president}$= [0.1, 2.4, 0.3], which represents the word "president" in a 3-dimensional space according to word2vec. In this space, words with similar meanings, such as "minister" and "president," are located in close proximity. However, the specific meaning of the value 2.4 in $w_{president}$ remains unclear. Consequently, it becomes difficult to answer questions regarding the significance of high and low values in the columns of W, as well as how to interpret the dimensions of word vectors. Prior work addressed interpretability in word embeddings through sparse and non-negative constraints on the embedding space [7-9] or post-processing pre-trained embeddings [6,10,11]. We propose a novel approach leveraging random matrix theory. We analyze eigenvectors of word embeddings derived from truncated SVD of the PPMI matrix [12-14]. We compare this to analyzing row and column spaces of Skip-Gram Negative Sampling (SGNS) used in word2vec training [15]. Based on the finding that SVD and SGNS factorize the same matrix [16], we posit that analyzing PPMI's principal eigenvectors can reveal information encoded in SGNS.

In recent times, numerous studies have presented similar findings regarding the semantic grouping of column values. Various algorithms have been proposed to train non-negative sparse interpretable word vectors [7–9, 17].

Our analysis identified a unique characteristic of the principal eigenvector (u1) that warrants further investigation. Unlike other eigenvectors, u1 exhibits a significant deviation from a normal distribution. This implies that the

information encoded within u1 does not follow the typical pattern observed in the remaining eigenvectors. Additionally, all elements of u1 possess negative, non-zero values. This stands in stark contrast to the potentially positive or negative values found in other eigenvectors. This unique behavior of u1 compels us to explore its potential role in the embedding space. It is tempting to speculate that u1 might capture a common bias that impacts all "words" within the corpus. This aligns with the findings presented in [19], which explored the influence of news events on stock prices. The consistently negative values within u1 might hold a connection to this phenomenon, potentially reflecting a general negative sentiment associated with news events. However, further investigation is necessary to solidify this hypothesis.

It is important to acknowledge prior efforts in achieving interpretability within the column space of word embeddings. For instance, the work presented in [6] proposed techniques for post-processing pre-trained word vectors by introducing non-negativity and sparsity constraints. While their approach demonstrably improved interpretability, the optional binarization of the vectors introduced a new challenge. The binary nature of the data makes it difficult to interpret the intensity or strength of the relationships between words compared to a real-valued representation.

Building upon this prior research, the work in [10] offered a solution that addressed the limitations of binarization. Their approach involved training a rotation matrix specifically designed to transform pre-trained word2vec and GloVe vectors. This method achieved interpretability without imposing the limitations of sparsity or binary constraints. Additionally, the research presented in [11] explored a similar path by post-training pre-trained word embeddings using k-sparse autoencoders, incorporating constraints akin to those used in [6]. While these efforts – alongside others such as [13-17] – have demonstrably achieved interpretability in the column space, as evidenced through word intrusion detection tests, they come with certain limitations. A common thread among these approaches is the reliance on either sparsity and non-negativity constraints or extensive post-processing steps. Furthermore, some of these methods dedicate less emphasis to analyzing and discussing the actual meanings associated with the interpretable dimensions, despite their stated goal of achieving interpretability.

In contrast, our research prioritizes exploring the implications of interpretability in the column space using conventional algorithms. We deliberately avoid introducing any additional constraints or post-processing steps. This allows us to leverage the inherent structure within the pre-trained word embeddings and potentially gain deeper insights into the underlying semantic relationships captured by the principal eigenvector (u1) and other interpretable dimensions. By eschewing additional processing steps, we aim to gain a clearer understanding of the actual meanings encoded within these interpretable dimensions without introducing potential biases or artifacts through external manipulations.

METHODS

For integrating insights from the Random Matrix Theory literature, we employ a novel approach to examine the eigenvectors of M^{PPMI}. Our analysis encompasses the investigation of eigenvector distributions, the computation of Inverse Participation Ratios (IPR) to gauge the proportion of noteworthy components, the assessment of structural sparsity, and the qualitative interpretation of these significant elements.

The eigenvector elements u^k are analyzed by comparing their empirical distribution with a normal distribution $N(\mu_{u^k}, \sigma^2_{u^k})$ in order to assess the normality of the eigenvectors. The parameters $\mu_{u^k}$and $\sigma^2_{u^k}$ correspond to the mean and variance of u^k.

Previous research [19] has demonstrated that eigenvectors that do not follow a Gaussian distribution exhibit significant correlations among stocks, as well as a global bias reflecting impactful news events that affect all stocks. These patterns will be further investigated in Section 5.1.

The Inverse Participation Ratio (IPR), represented as I^k, serves as a measure to quantify the inverse proportion of significant elements within the eigenvector u^k [19–21].

where u_i^k is the i-th element of u^k . The intuition of IPR can be illustrated with two extreme cases. The first one is when all elements of u^k are equal to the same value $1/\sqrt{|v|}$ which results in I^k is equal to $1/|v|$ with reciprocal $\frac{1}{I^k}$ being $|v|$. As a result, this makes all elements contribute in a similar way. The other case is when one element is used as a one-hot vector with only one element as one, and the rest as zero. In this case u^k will have an IPR value of one and the sane goes for reciprocal.

Visualization of the most significant elements of the top eigenvector

Considering u^k, $v^k \in \mathbb{R}|v|$, we have the ability to associate each index of the vectors with a word in the vocabulary V. Consequently, we explore the dimensions and their corresponding indices (or words) that possess the highest absolute values, aiming to identify semantic consistency. Previous studies employing similar methodologies with financial data have successfully clustered stocks from the same industries or proximate regions [19]. Likewise, in the context of genetic

data, this approach has unveiled crucial co-evolving genes within gene co-expression networks [20].

RESULTS

Training

We utilized the English Wikipedia [21] dump that was processed using Matt Mahoney's Perl script, a tool that has been referenced in previous studies. By eliminating extraneous non-alphanumeric characters like XML tags, the dataset size was significantly decreased from about 66GB to 25GB, encompassing roughly 3.4 billion tokens. The vocabulary size is estimated to be around 346,000, focusing solely on words that appear at least 100 times.

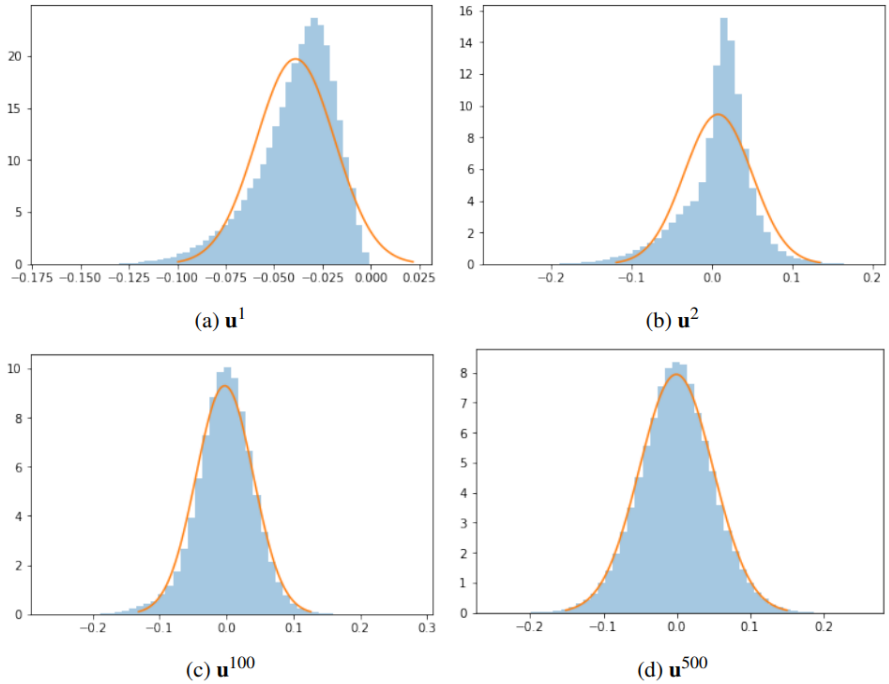


Figure 1. The eigenvector distributions of u^1, u^2, u^{100}, and u^{500} (where u^1 represents the largest eigenvector) are depicted in the graph. The solid curves in the graph represent Gaussian distributions.

Distribution of Eigenvector Elements

From the data presented in Figure 1, it is evident that eigenvectors associated with the larger eigenvalues, such as u^1 or u^2, exhibit a noticeable departure from a Gaussian distribution. Similarly, u^{100} and u^{500} also show deviations, albeit to a lesser extent. This observation indicates that the eigenvectors possess non-random characteristics and contain significant correlations. Such a pattern is expected since these vectors are the principal eigenvectors.

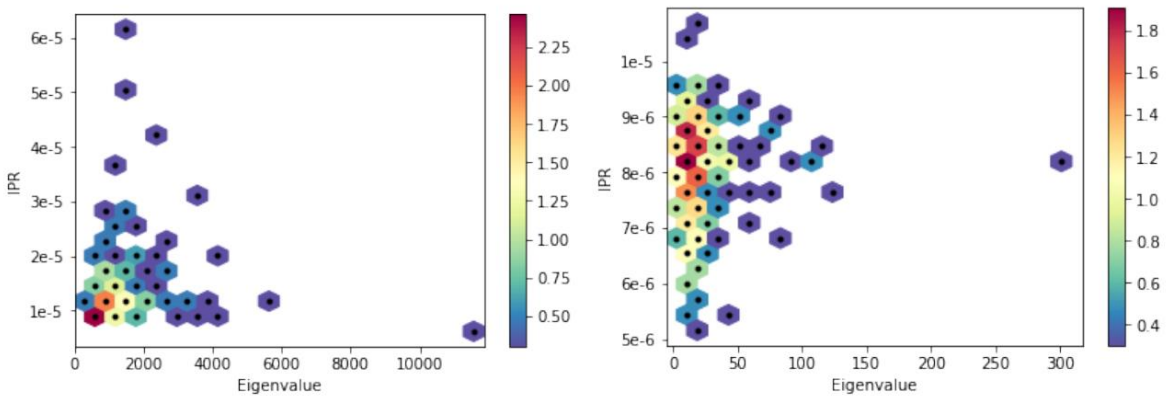
Our analysis revealed a unique characteristic of the principal eigenvector (u1). Unlike other eigenvectors, u1 exhibits significant deviation from a normal distribution and contains only negative, non-zero values. This suggests u1 might capture a global bias affecting all "words" potentially linked to the influence of news events on stock prices, as explored in [19].

The Inverse Participation Ratio

Figure 2 depicts the Inverse Participation Ratio (IPR) of u^k in relation to the eigenvalue λ^k, and similarly for v^k. The visualization highlights that the eigenvectors of W^{SVD} exhibit IPR values around 10 times greater than those of W^{SGNS}, indicating a significantly sparser nature of vectors in W^{SVD}.

From the data presented in Figure 2a, it is evident that the eigenvector with the highest magnitude has the lowest Inverse Participation Ratio (IPR) value of 0.000006. Furthermore, when the reciprocal of I^k is divided by the absolute value of V, it results in a value of 48%. In contrast, the largest I^k yields a value of approximately 4.7%. The average value of the reciprocal of I^k divided by the absolute value of v, across all eigenvectors, is 27.5%. This suggests the presence of some sparse structure within the eigenvectors of W^{SVD}.

On the other hand, Figure 2b illustrates that the mean value for v^k is approximately 36%. This indicates that the column vectors of W^{SGNS} are generally denser and less structured compared to W^{SVD}. The observed difference in structural sparsity between the two methods motivates us to conduct a thorough analysis of the eigenvectors of W^{SVD}.



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

Figure 2. The concentration of points increases as the dots become redder, as indicated by the inverse participation ratios.

DISCUSSION

Based on the findings from the preceding sections, we proceed to analyze the principal components of the eigenvectors by arranging their absolute values in a descending order. The outcomes are presented in Table 1, revealing intriguing patterns as the prominent dimensions or the associated "words" of each eigenvector tend to create semantically or syntactically cohesive clusters. For example, u^{14} clusters French terms together, while u^{121} displays words related to baseball. Some terms within u^{121} may initially appear unrelated to baseball. Nevertheless, "buehrle" is a baseball player, "rbis" represents "Run Batted Ins", and "astros" denotes a baseball team based in Houston. Conversely, the terms grouped in u^{1}, the most significant eigenvector, could elucidate the bias discussed in Training Section. The noteworthy elements appear to be potent transitional words frequently utilized for dramatic impact, such as "importantly" or "crucially". Clearly, these terms heighten the intensity of the context.

Table 1. The primary eigenvectors (dimensions with the greatest magnitudes) of WSVD organize into semantically meaningful clusters. Eigenvectors u^{14} and u^{121} exhibit high IPR values, whereas the rest represent eigenvectors associated with the largest eigenvalues.

$\mathbf{u}^{1}$	$\mathbf{u}^{4}$	$\mathbf{u}^{7}$	$\mathbf{u}^{8}$	$\mathbf{u}^{14}$	$\mathbf{u}^{121}$
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

CONCLUSION

In this study, we examined the eigenvectors, specifically the column space, of the word embeddings derived from the Singular Value Decomposition of the PPMI matrix. Our analysis unveiled that the prominent components of the

eigenvectors cluster together semantically, enabling us to perceive each word vector as a comprehensible feature vector comprised of semantic clusters. These findings hold potential for aiding in error detection in subsequent NLP assignments, or selecting valuable feature dimensions to streamline the development of compact and effective task-specific embeddings. Subsequent research will continue to explore the application of interpretability in practical contexts.

Acknowledgement
Authors declare any financial support or relationships that may pose conflict of interest in the covering letter submitted with the manuscript.

Conflict of interest
The authors declare that they have no conflict of interest associated with this manuscript.

REFERENCES

1. HARRIS Z. Distributional structure. 1954.
2. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013;26.
3. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 Oct (pp. 1532-1543).
4. KIM, Hee-Cheol. Convolutional neural network for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) Association for Computational Linguistics.2014. p. 1746-1751.
5. Faruqui M, Tsvetkov Y, Yogatama D, Dyer C, Smith N. Sparse overcomplete word vector representations. arXiv preprint arXiv:1506.02004. 2015 Jun 5.
6. Murphy B, Talukdar PP, Mitchell T. Learning effective and interpretable semantic models using non-negative sparse embedding. InInternational Conference on Computational Linguistics (COLING 2012), Mumbai, India 2012 Dec (pp. 1933-1949). Association for Computational Linguistics.
7. Luo H, Liu Z, Luan H, Sun M. Online learning of interpretable word embeddings. InProceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 2015 Sep (pp. 1687-1692).
8. Sun F, Guo J, Lan Y, Xu J, Cheng X. Sparse word embeddings using l1 regularized online learning. InTwenty-Fifth International Joint Conference on Artificial Intelligence 2016 Jul 9.
9. Park, Sungjoon; Bak, Jinyeong; OH, Alice. Rotated word vector representations and their interpretability. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. p. 401-411.
10. Subramanian A, Pruthi D, Jhamtani H, Berg-Kirkpatrick T, Hovy E. Spine: Sparse interpretable neural embeddings. InProceedings of the AAAI conference on artificial intelligence 2018 Apr 26 (Vol. 32, No. 1).
11. Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika. 1936 Sep;1(3):211-8.
12. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American society for information science. 1990 Sep;41(6):391-407.
13. Church K, Hanks P. Word association norms, mutual information, and lexicography. Computational linguistics. 1990;16(1):22-9.
14. Mikolov T. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
15. Kober T, Weeds J, Reffin J, Weir D. Improving sparse word representations with distributional inference for semantic composition. arXiv preprint arXiv:1608.06794. 2016 Aug 24.
16. Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Transactions of the association for computational linguistics. 2015 May 1;3:211-25.
17. Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Guhr T, Stanley HE. Random matrix approach to cross correlations in financial data. Physical Review E. 2002 Jun 27;65(6):066126.
18. Jalan S, Solymosi N, Vattay G, Li B. Random matrix analysis of localization properties of gene coexpression network. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics. 2010 Apr;81(4):046118.

https://journal.utripoli.edu.ly/index.php/Alqalam/index eISSN 2707-7179

تحليل الخصائص الدلالية لتضمينات الكلمات باستخدام المتجهات الذاتية

بسمة عبدالرازق*, حنان عطية الله

قسم الرياضيات, كلية العلوم, جامعة عمر المختار, البيضاء , ليبيا

المستخلص

لقد أثبتت متجهات الكلمات الكثيفة فعاليتها في العديد من مهام معالجة اللغة الطبيعية اللاحقة في السنوات الأخيرة. ومع ذلك، لا تزال قابلية تفسير أبعاد هذه التضمينات تشكل تحديًا. في هذه الورقة، نستكشف كيف يمكن للمتجهات الذاتية أن تكشف عن خصائص دلالية مختلفة تم التقاطها بواسطة نماذج تضمين الكلمات. وبالتالي، نقوم بتدريب تضمينات الكلمات (على سبيل المثال، Word2Vec) على مجموعة ويكيبيديا الإنجليزية لتحليل المتجهات الذاتية العليا، وتحديد الخصائص الدلالية المحددة (على سبيل المثال، المشاعر، والشكلية) المرتبطة بكل منها، واستكشاف كيفية ترميز هذه الخصائص في مساحة التضمين. ناقشت هذه الورقة أيضًا القيود والفوائد المحتملة لهذا النهج مقارنة بالطرق الأخرى لتحليل تضمينات الكلمات.

الكلمات المفتاحية: متجهات الكلمات، تضمينات الكلمات، مساحة التضمين، المتجهات الذاتية.