Getting in Shape: Word Embedding SubSpaces

Tianyuan Zhou[1] , **João Sedoc**[2] and **Jordan Rodu**[1*]
[1]Department of Statistics, University of Virginia
[2]Department of Computer and Information Science, University of Pennsylvania
tz8hu@virginia.edu, joao@cis.upenn.edu, jsr6q@virginia.edu

Abstract

Many tasks in natural language processing require the alignment of word embeddings. Embedding alignment relies on the geometric properties of the manifold of word vectors. This paper focuses on supervised linear alignment and studies the relationship between the shape of the target embedding. We assess the performance of aligned word vectors on semantic similarity tasks and find that the isotropy of the target embedding is critical to the alignment. Furthermore, aligning with an isotropic noise can deliver satisfactory results. We provide a theoretical framework and guarantees which aid in the understanding of empirical results.

1 Introduction

Mono-lingual and multi-lingual alignment of word embeddings is important for domain adaptation, word embedding assessment, and machine translation [Ben-David *et al.*, 2007; Blitzer *et al.*, 2011; Tsvetkov *et al.*, 2015; Lample *et al.*, 2018]. Fundamentally, this is a subspace alignment problem which has seen much interest in machine learning communities, including computer vision and nature language processing (NLP) [Fernando *et al.*, 2013; Xing *et al.*, 2015; Wang and Mahadevan, 2013; Lample *et al.*, 2018; Mikolov *et al.*, 2013a] and can be either supervised or unsupervised, depending on the setting.

Simultaneously, some work has focused on uncovering the structure of word embedding manifolds [Mimno and Thompson, 2017a; Hartmann *et al.*, 2018; Shin *et al.*, 2018; Mu *et al.*, 2017], where in this paper we focus their isotropy/anisotropy. Word embeddings are known to not be isotropic [Andreas and Klein, 2015], but Arora [2015] argue that isotropic word embeddings mitigate the effect of approximation error. Indeed, recent work in post-processing word embeddings has shown that increasing isotropy increases semantic task performance [Mu *et al.*, 2017; Liu *et al.*, 2019b; Liu *et al.*, 2019a].

Despite a large body of work in each of the two areas, the link between subspace alignment and manifold isotropy has not been fully explored. Given that word embeddings contain some form of "meaning", presumably common to all word embeddings, different representations should align to some degree. But alignment necessarily prioritizes resolving disparity in larger singular values, which could be problematic for two word embeddings that encode information differently across their spectrums. When two word embeddings represent information similarly, they can be successfully aligned. For instance, Artetxe [2016] shows orthogonal transformations for alignment are superior for retaining performance on analogy tasks. However, orthogonal transformations do not allow for the alignment of more disparate methods such as distributional and non-distributional word embeddings.

In this paper, we present a theoretical framework for understanding the alignment between word embeddings- when they can work, and when they might fail. Our theoretical results show that the singular value structure of the source embeddings is completely discarded, and when information is encoded differently in the source and target embeddings, the distortion of the spectrum from source to aligned embeddings could drastically effect downstream results.

2 Theoretical Results

In this section we provide some theoretical underpinnings for the phenomena observed in our experiments. Lemma 1 shows that when a source representation is aligned to a target representation using a linear transformation, the column space of the aligned representation is determined by the column space of the source, but the singular value structure of the source is entirely discarded. Theorem 1 guarantees the existence of a lower-bounded singular value. Since the summation of the singular values is upper-bounded by the target embedding, if the lower bound of the largest singular value is relatively large in comparison, then the singular value structure will have high eccentricity.

Finally, proposition 1 shows that the correlation of measurements between words (Euclidean distance between vectors or the cosine of their angles) in a vector cloud, and those measurements after the vector cloud has been stretched by unequal amounts in different directions, is small. Combined with the results of theorem 1, this implies that alignment can greatly

*Corresponding Author
Link of Supplementary Materials and Source codes: https://github.com/NoahZhouTianyuan/ConceptorOnNondistEmbedding

impact the performance of aligned word representations on semantic similarity tasks. We expand on this below.

We include the proof of the theorem here, but defer the proof of the lemma and the proposition to the supplementary material. We also provide additional theorems in the supplementary material that, while not critical for the main focus of this work, provide justification for some minor experimental results and contribute more broadly to our fundamental theoretical understanding of word representation alignment.

We begin by stating our word representation alignment model.

Alignment Model: Let $Y \in \mathbb{R}^{n \times p}$ be the target word representations, and $X \in \mathbb{R}^{n \times p}$ the source representations. To align X to Y, we seek a linear mapping W such that $Y = XW + \epsilon$. The least squares solution for W yields $\widehat{Y} = XW^* = X(X^TX)^{-1}X^TY = U_XU_X^TY$.

Lemma 1. *Let U_X and U_Y be the left singular vectors of X and Y respectively, and Σ_Y a diagonal matrix containing the singular values of Y, then $\sigma(\widehat{Y}) = \sigma(U_X^TU_Y\Sigma_Y)$.*

Corollary 1. *If $U_X = U_Y$, then $\sigma(\hat{Y}) = \sigma(Y)$.*

For the following theorem, define $\sigma_i(M)$ to be the i^{th} singular value of the matrix M, and U_{Mi} to be the singular vector associated with the i^{th} singular value of M.

Theorem 1. *Suppose $X, Y \in \mathbb{R}^{n \times p}$, and let $\widehat{Y} = XW^*$ be the least squares solution of $Y = XW + \epsilon$. For singular value $\sigma_i(Y)$ and corresponding left singular vector U_{yi}, let $c_i = ||U_X^\top U_{yi}||_2$, then there exists a singular value in $\widehat{Y}$ at least as large as $c_i\sigma_i(Y)$.*

Proof. *By Lemma 1, $\sigma(\hat{Y}) = \sigma(U_X^TU_Y\Sigma_Y)$. Recalling the definition of singular values, $\sigma_1(U_X^TU_Y\Sigma_Y) = \max_{||u||_2=1,||v||_2=1} u^TU_X^TU_Y\Sigma_Y v$ where u, v are unit vectors. Define $e_i = (0, 0, \cdots, 0, 1, 0, \cdots, 0)^T$ where e_i is a unit vector whose elements are 0 except the i^{th} element which is 1. Then for any $1 \leq i \leq p$, $\sigma_1(U_X^\top U_Y\Sigma_Y) \geq \max_{||u||_2=1} u^TU_X^TU_Y\Sigma_Y e_i = \max_{||u||_2=1} u^\top U_X^\top\sigma_i(Y)U_{yi} = \sigma_i(Y)||U_X^\top U_{yi}||_2 = c_i\sigma_i(Y)$. Therefore, the largest singular value must be greater than or equal to $c_i\sigma_i(Y)$, for all i.*

Proposition 1. *Suppose $X \in \mathbb{R}^p$ is a random vector with all entries distributed i.i.d with mean zero and a bounded fourth moment. Let S be a matrix with $\sqrt{s_1}, \sqrt{s_2}, \cdots, \sqrt{s_p})$ along the diagonal and 0 everywhere else. Then for realizations X_i and X_j of X, we have the following two results:*

$$corr(||X_i - X_j||_2^2, ||X_iS - X_jS||_2^2) = \frac{(s_1 + s_2 + \cdots + s_p)/\sqrt{p}}{\sqrt{s_1^2 + s_2^2 + \cdots + s_p^2}}$$

and

$$corr(< X_i, X_j >, < X_iS, X_jS >) = \frac{(s_1^2 + s_2^2 + \cdots + s_p^2)/\sqrt{p}}{\sqrt{s_1^4 + s_2^4 + \cdots + s_p^4}}$$

Further, for a subset $I \subset \{1, \ldots, p\}$ define $X_{i,I}$ to be the vector X_i with indices not in I set to 0. Then

$$corr(||X_iS - X_jS||_2^2, ||X_{i,I} - X_{j,I}||_2^2) = \frac{\sum_{i \in I} s_i/\sqrt{|I|}}{\sqrt{s_1^2 + s_2^2 + \cdots + s_p^2}}$$

To see the implications of this proposition, consider the alignment process loosely as 1) establishing the directions of the aligned singular vectors (which could be by an orthogonal rotation of the source singular vectors, or through linear combinations of the source embeddings if the information content is not encoded similarly in the source and target embeddings), 2) stretching the singular vectors according to the singular values of the source embeddings (or functions of the singular values that distribute their information appropriately if the transformation is not orthogonal), and 3) adjusting the stretched singular vectors through the S matrix of proposition 1 according to the spectrum of the target embeddings (see lemma 1). For two word embeddings that encode information similarly (say two distributional word embeddings) the entries of the S matrix will all be roughly equal. However, for two word embeddings that do not, it is likely that significant adjustment will be required, with some entries in the S large to stretch the singular vectors, and some small to shrink them. The results of our experiments support this idea.

3 Empirical Results

In this section, we show some empirical results of word representation alignment. Our key finding suggests that isotropy is important to successful alignment. In fact, aligning with isotropic noise can even yield satisfactory intrinsic evaluation results.

Conceptor Negation (CN). It is worth noting that the geometry of the distributional word embedding has been studied carefully [Mimno and Thompson, 2017b]. Mimno et al. [2017b] note that for word2vec word embedding point clouds are concentrated within a narrow cone, which may lead to bias. ABTT [Mu *et al.*, 2017] and conceptor negation [Liu *et al.*, 2019b] are two methods used to correct this bias by damping the larger singular values. Hence we suggest that conceptor negation should be used post alignment in order to control the eccentricity of the resulting aligned representation.

3.1 Experimental Setup

We perform multiple experiments using distributional word representations (each 300-dimensional) including word2vec [Mikolov *et al.*, 2013b] (Google News), GloVe [Pennington *et al.*, 2014] (840 billion Common Crawl) and FastText [Bojanowski *et al.*, 2017] (Common Crawl without subword), as our source embeddings, and align them through linear regression to various target representations. We then test the aligned word vectors on seven similarity tasks [Faruqui and Dyer, 2014], and in some cases an additional three concept categorization tasks as supplement. Our target representations are (1) other distributional word representations; (2) word representations such as low-rank non-distributional word vectors

with singular values [Faruqui and Dyer, 2015], LSA [Landauer *et al.*, 1998] (on British National Corpus [Consortium, 2007]), Eigenwords [Dhillon *et al.*, 2015] and dependency word embedding [Levy and Goldberg, 2014a]; (3) The same word representations in (2) but with the labels permuted; and (4) designed noise, including isotropic Gaussian noise and rank-1 noise. We also ran experiments in which we performed conceptor negation on the target, source, or aligned word representation, and compared these results to the non-conceptored versions.

Note that the shapes of the manifolds of the different distributional word representations are quite similar and the alignment of distributional source representations to distributional target representations has a high R^2. In contrast, the shapes of the manifolds of the word representations in (2) are quite different than in distributional word representations, and as a consequence the alignment has a relatively low R^2. The permuted embeddings disassociate the labels with the representations, thus breaking the supervised nature of alignment. Finally, the designed noise simulates two extreme cases, one purely isotropic target representation and one highly non-isotropic.

Alignment Order for Display of Semantic Similarity Results

To better display and compare the results of our alignment experiments on semantic similarity tasks, we establish a sequence of alignments. For aligning two distributional word representations (call them representation 1 and representation 2), the alignment sequence is (in parentheses are example labels of taking word2vec as representation 1 and GloVe as representation 2): (1) representation 1 (e.g. W2V), (2) conceptor negated representation 1 (e.g. W2V+CN), (3) representation 1 aligned to conceptor negated representation 2 (e.g. W2V by GV+CN), (4) representation 1 aligned to representation 2 (e.g. W2V by GV), (5) representation 2 aligned to representation 1 (e.g. GV by W2V), (6) representation 2 aligned to conceptor negated representation 1 (e.g. GV by W2V+CN), (7) conceptor negated representation 2 (e.g. GV+CN), (8) representation 2 (e.g. GV). Results for distributional sources and targets are shown in Fig 1.

For a distributional source aligned with an 'other' target, the sequence is (in parentheses are the labels shown in figures): (1) Source representation (Source Embedding), (2) source aligned with permuted and conceptor negated target (Aligned by permuted+CN Target), (3) source aligned with permuted target (Aligned by permuted Target), (4) source aligned with target (Aligned by Target), (5) target representation (Target Embedding). For example, the results for the FastText source and non-distributional target representations are shown in Fig 4.

3.2 Alignment Results

Distributional Source and Target Representations

Distributional word vector representations model similar information [Levy and Goldberg, 2014b; Pennington *et al.*, 2014] implying that the manifold shapes are similar across representations. We show that aligning with conceptor negated target vectors leads to improved semantic similarity task scores. From *Corollary* 1, we have that, when the subspaces of the source and target embeddings are nearly identical, the singular value information of the target representations dictate the information of the aligned word representations, which predicts that conceptor negating the target representation yields good results. The results are shown in Fig 1.

Distributional Source with Other Word Representations

Distributional source with non-distributional targets. Non-distributional word representations typically encode lexical information, and thus the shape of the representation manifold is quite different from that of distributional representations, confirmed by their differences in performance as seen in Fig 3. Alignment of distributional word representations to non-distributional word representations highlights another important consequence of *Lemma* 1. Namely, both the singular values of the target representations, and the degree of overlap in the subspaces spanned by the source and target representations, matter in determining the singular value structure of the aligned representations. Importantly, the distributional and non-distributional subspaces seem to overlap only in directions with high target singular values- Fig 2 shows that singular vectors corresponding to large singular values have much higher R^2. From *Theorem* 1 then, the largest singular value of the aligned representation is lower-bounded, particularly by $||U_X^\top U_{y1}||_2 \sigma_1(Y)$ as both $\sigma_1(Y)$ and $||U_X^\top U_{y1}||_2$ are largest among $\sigma_i(Y)$s and $||U_X^\top U_{yi}||_2$s respectively, resulting in highly eccentric aligned word representations. Also, the overall R^2 when the target embedding is not conceptor-negated is higher because these singular vectors contribute more than when the target embedding is conceptor-negated. This suggests that the best way to combine the information from distributional and non-distributional word vectors should be through concatenation instead of alignment. The sequential comparisons (Fig 4) generally show that alignment with the 'permuted target' ('Aligned by permuted Target' in figures) or the unchanged target ('Aligned by Target' in figures) tend to give the worst results, while aligning with 'permuted + conceptor negated target' ('Aligned by permuted+CN Target' in figures) typically performs much better. A major difference between the source representation and the aligned representation is the singular value structure, as the singular vector structure of the source is lost. This explains why aligning with the 'permuted + conceptor negated target' does relatively well. The information from the target singular values is tempered, making the aligned vectors more isotropic. In contrast, aligning with 'permuted target' or the unchanged target transmit full information of the target singular values. Target singular value information seems most detrimental when the source and target representations cannot be aligned well. Note that although aligning with 'permute + conceptor negated target' performs well on semantic similarity tasks, the R^2 is nearly 0.

Distributional source with other targets (except non-distributional). For distributional source representations aligned with other target representations (Figs 5 - 10), the sequential comparison trends are similar: aligning the distributional word vectors with permuted target representations yields the worst results in most cases, while aligning with the unchanged target performs worse than aligning with permuted and conceptor negated target. In some cases, aligning with

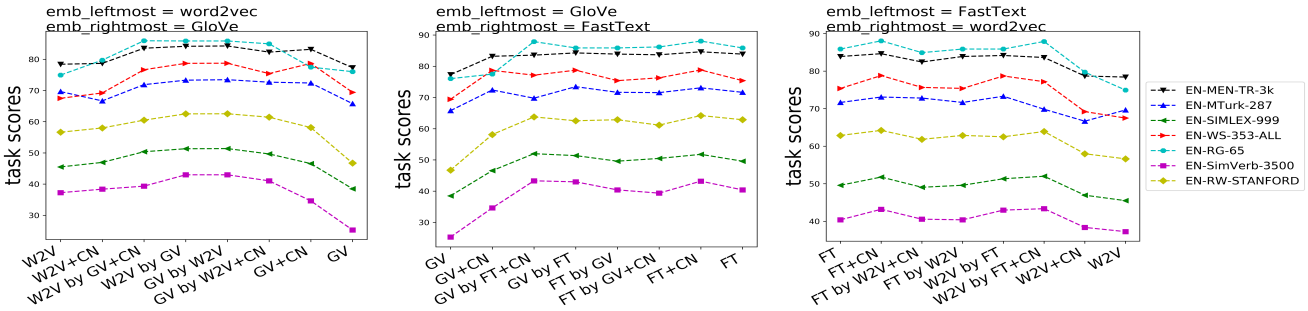

Figure 1: Comparison of similarity task scores between distributional word vectors, the conceptor negated vectors, and aligned vectors. Here we compare between word2vec (W2V), GloVe (GV) and FastText (FT).

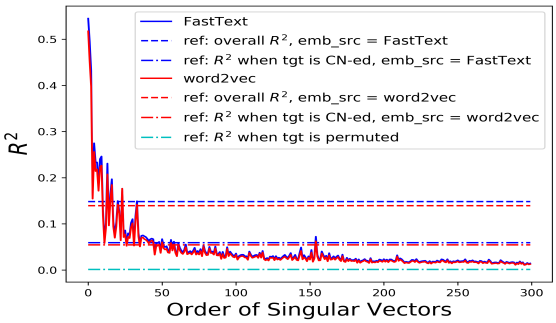

Figure 2: R^2 of each singular vectors of non-distributional word vectors fitted by distributional word vectors. X-axis is the order of singular vectors, starting from the largest singular value to the smallest. For example, $X = 1$ means the singular vector corresponding to the largest singular value, and $X = 2$ means the singular vector corresponding to the second largest singular value. The results when the source embedding is GloVe is nearly identical as word2vec so we do not show it otherwise it will be covered. The imaginary reference lines show the overall R^2 of the alignments, while the dot imaginary lines show the overall R^2 of the alignments when the target word vectors are conceptor negated.

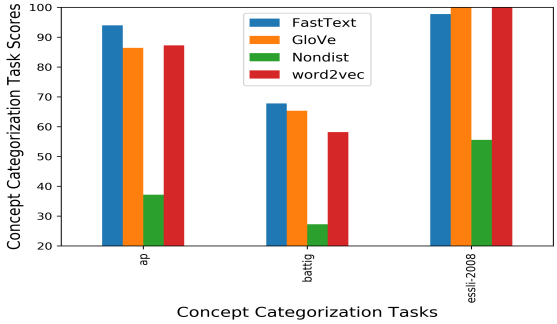

Figure 3: Scores of concept categorization tasks of distributional and non-distributional word vectors. The non-distributional word vectors performed worse than distributional word vectors on all three tasks. This illustrates why distributional and non-distributional word vectors can hardly be aligned, as their magnitude information are different.

a permuted and conceptor negated target yields better results than the source embedding, especially GloVe, though the R^2 is very low. Results of alignments with dependency CBOW and GloVe (500 dimensions) are included in the supplementary material.

Distributional Source with Designed Noise Target

To further understand the scope of and issues with alignment, we align the distributional source representations with two types of designed noise targets, isotropic Gaussian noise and rank-1 noise. The results (Fig [11]), show that aligning with Gaussian noise only minimally decreases the performance from the performance of the source embedding, while aligning with rank-1 noise destroys information across the board. This result is a consequence of *Proposition* 1.

4 Conclusion

Despite the success of word representations in NLP tasks, relatively little is known about their theoretical properties. In this paper we try to aid understanding through studying the properties of word representation alignment. We show through theory and experiment that the target representations drive the information content of the singular values of the aligned representations, while the source representations preserve the subspace. In theory, a carefully designed experiment with carefully constructed tasks could help tease apart what aspects of the word representations encode which characteristics of words. We lay this groundwork here.

Further, one must take care when performing word representation alignment. While there are demonstrated benefits to alignment, they do not uniformly apply to semantic similarity tasks. Our theoretical framework provides guidance as to why alignment might work in some cases but not in others. For instance, our theory and results provide justification for why concatenating distributional and non-distributional word representations is preferable to alignment.

Establishing a theoretical foundation for understanding word representations will provide impetus for improved performance of word representations in NLP tasks. In this paper, we lay the groundwork for understanding alignment which, in addition to allowing for the integration of information across word representations, can provide a unique lens into how word representations encode information.

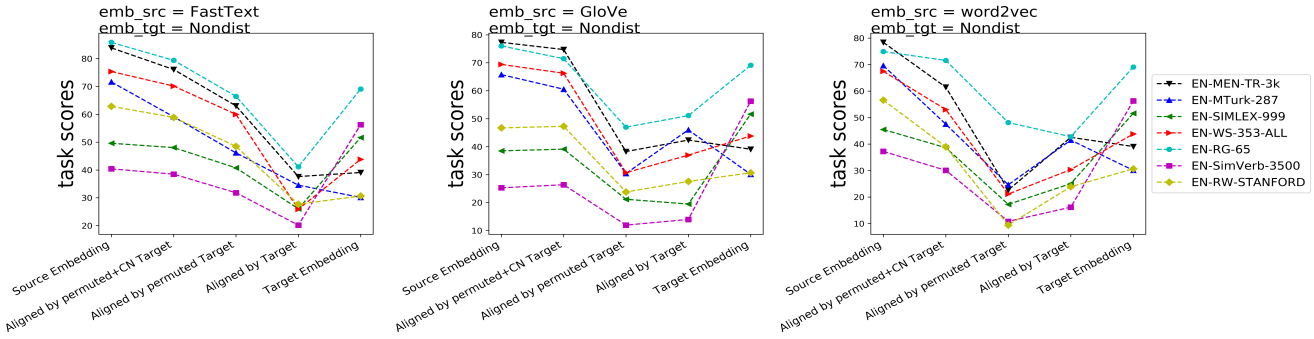

Figure 4: Sequential Comparison of FastText/GloVe/word2vec Aligned with Non-distributional Word Vectors

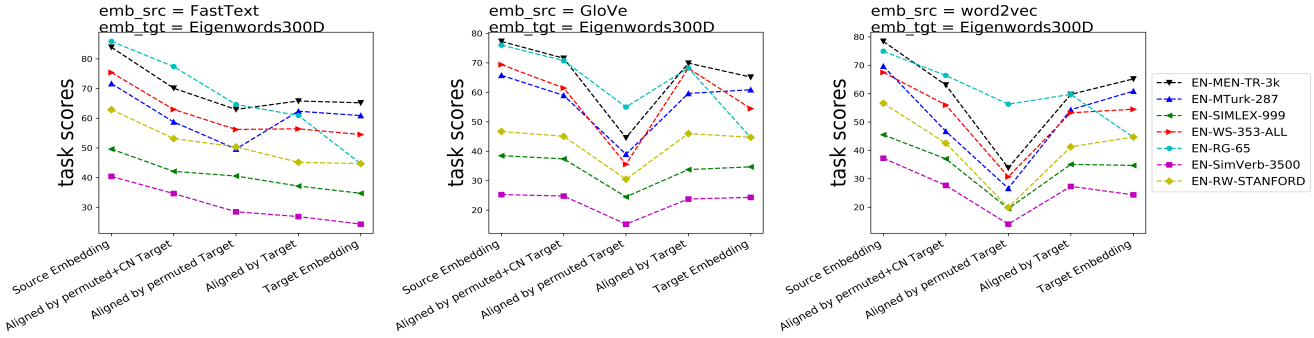

Figure 5: Sequential Comparison of FastText/GloVe/word2vec Aligned with Eigenwords

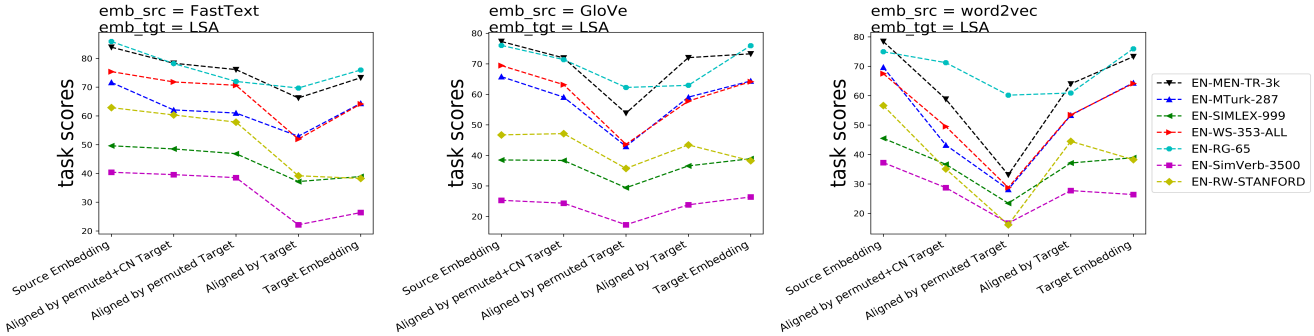

Figure 6: Sequential Comparison of FastText/GloVe/word2vec Aligned with LSA

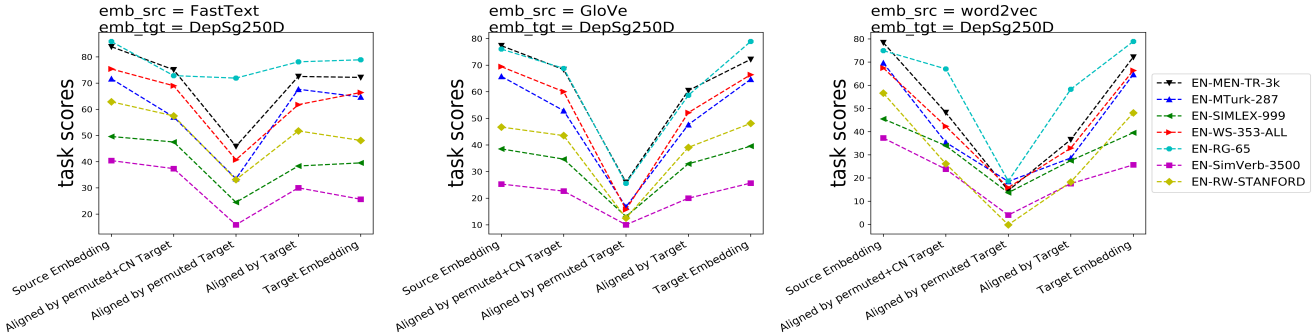

Figure 7: Sequential Comparison of FastText/GloVe/word2vec Aligned with Dependency Word Embedding (Skip-Gram 250D)

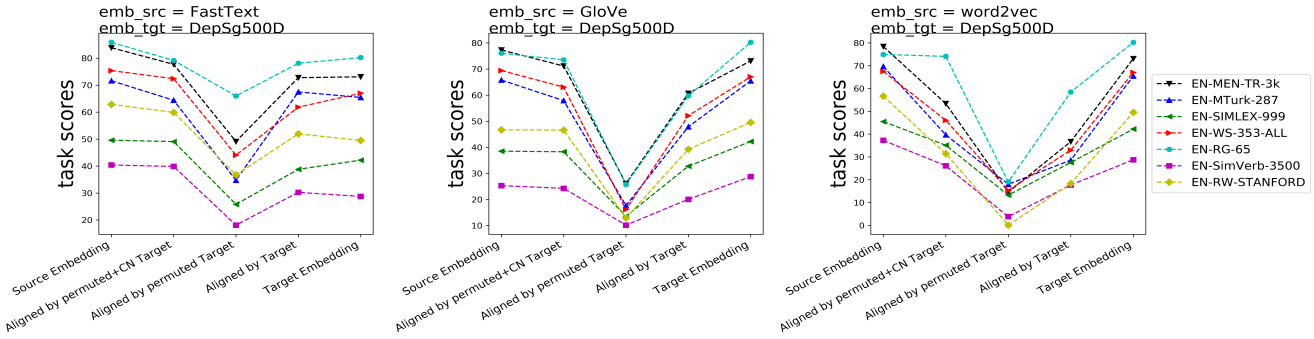

Figure 8: Sequential Comparison of FastText/GloVe/word2vec Aligned with Dependency Word Embedding (Skip-Gram 500D)

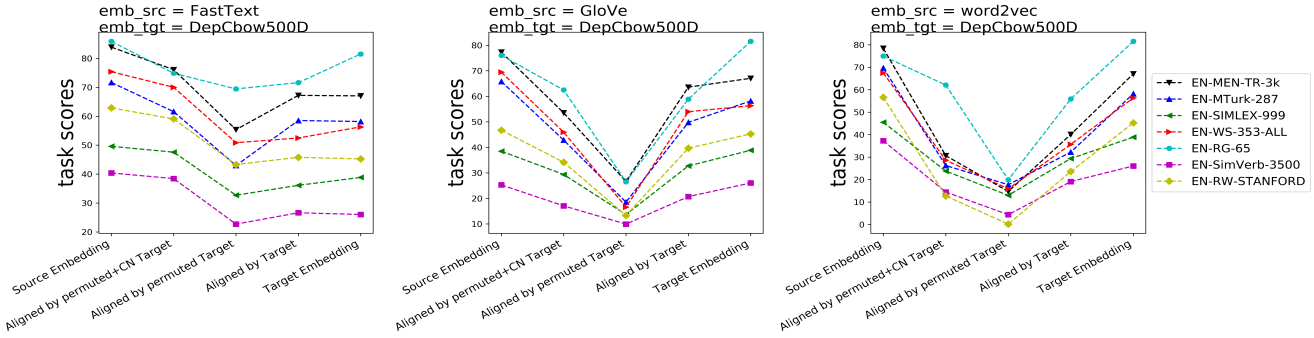

Figure 9: Sequential Comparison of FastText/GloVe/word2vec Aligned with Dependency Word Embedding (CBOW 500D)

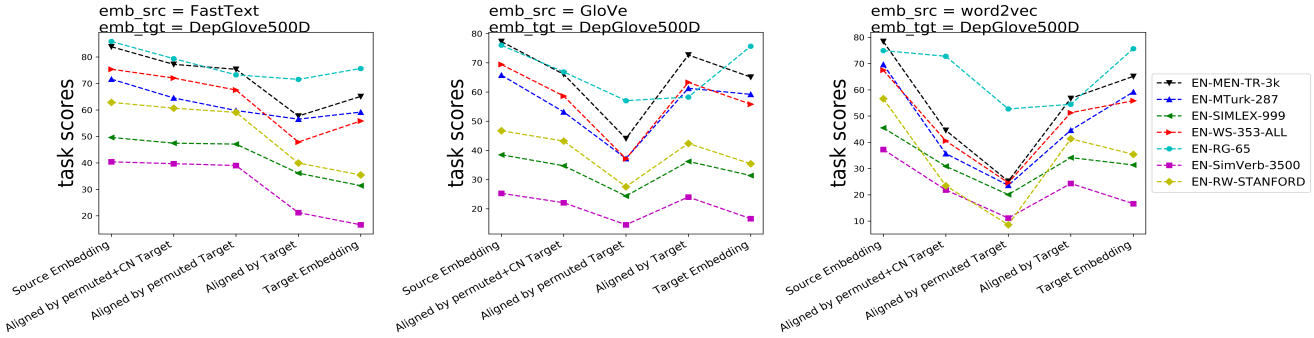

Figure 10: Sequential Comparison of FastText/GloVe/word2vec Aligned with Dependency Word Embedding (GloVe 500D)

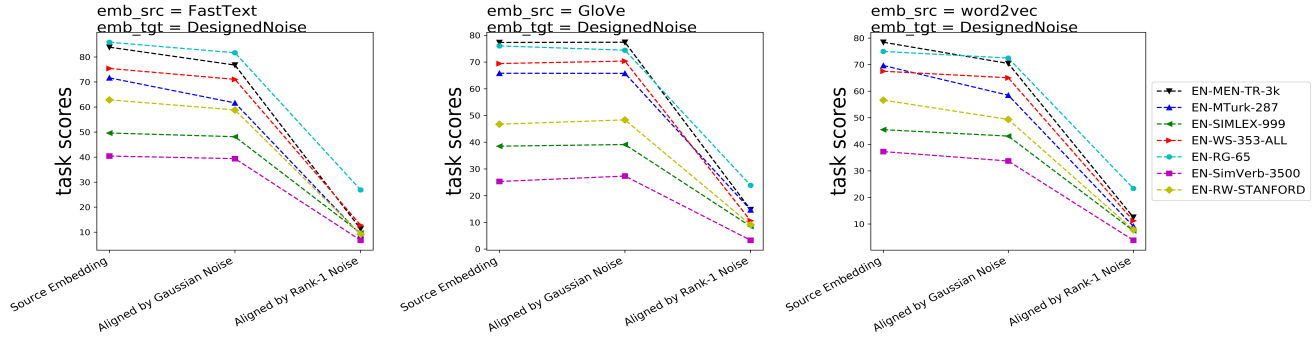

Figure 11: Similarity task scores of FastText/GloVe/word2vec Aligned with Designed Noise

References

[Andreas and Klein, 2015] Jacob Andreas and Dan Klein. When and why are log-linear models self-normalizing? In *Proceedings of NAACL HLT*, pages 244–249, 2015.

[Arora *et al.*, 2015] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Rand-walk: A latent variable model approach to word embeddings. *arXiv preprint arXiv:1502.03520*, 2015.

[Artetxe *et al.*, 2016] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*, pages 2289–2294, 2016.

[Ben-David *et al.*, 2007] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144, 2007.

[Blitzer *et al.*, 2011] John Blitzer, Sham Kakade, and Dean Foster. Domain adaptation with coupled subspaces. In *Proceedings of AISTAT*, pages 173–181, 2011.

[Bojanowski *et al.*, 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[Consortium, 2007] BNC Consortium. British national corpus, version 3 BNC XML edition, 2007.

[Dhillon *et al.*, 2015] Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. Eigenwords: Spectral word embeddings. *JMLR*, 16(1):3035–3078, 2015.

[Faruqui and Dyer, 2014] Manaal Faruqui and Chris Dyer. Community evaluation and exchange of word vectors at wordvectors. org. In *Proceedings of ACL: System Demonstrations*, pages 19–24, 2014.

[Faruqui and Dyer, 2015] Manaal Faruqui and Chris Dyer. Non-distributional word vector representations. In *Proceedings of ACL*, pages 464–469, Beijing, China, July 2015. Association for Computational Linguistics.

[Fernando *et al.*, 2013] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

[Hartmann *et al.*, 2018] Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. Why is unsupervised alignment of English embeddings from different algorithms so hard? In *Proceedings of EMNLP*, pages 582–586, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Lample *et al.*, 2018] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*, 2018.

[Landauer *et al.*, 1998] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[Levy and Goldberg, 2014a] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of ACL*, volume 2, pages 302–308, 2014.

[Levy and Goldberg, 2014b] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *NIPS*, pages 2177–2185, 2014.

[Liu *et al.*, 2019a] Tianlin Liu, Lyle Ungar, and João Sedoc. Continual learning for sentence representations using conceptors. In *Proceedings of NAACL HLT*, 2019.

[Liu *et al.*, 2019b] Tianlin Liu, Lyle Ungar, and João Sedoc. Unsupervised post-processing of word vectors via conceptor negation. In *Proceedings of AAAI*, 2019.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[Mimno and Thompson, 2017a] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Proceedings of EMNLP*, pages 2873–2878, 2017.

[Mimno and Thompson, 2017b] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Proceedings of EMNLP*, pages 2873–2878, 2017.

[Mu *et al.*, 2017] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective post-processing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, 2014.

[Shin *et al.*, 2018] Jamin Shin, Andrea Madotto, and Pascale Fung. Interpreting word embeddings with eigenvector analysis. *https://openreview.net/pdf?id=rJfJiR5ooX*, 2018.

[Tsvetkov *et al.*, 2015] Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*, pages 2049–2054, 2015.

[Wang and Mahadevan, 2013] Chang Wang and Sridhar Mahadevan. Manifold alignment preserving global geometry. In *IJCAI*, pages 1743–1749, 2013.

[Xing *et al.*, 2015] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the NAACL HLT*, pages 1006–1011, 2015.