ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Geometric Probing of Word Vectors

Madina Babazhanova Maxat Tezekbayev
Zhenisbek Assylbekov

School of Sciences and Humanities — Nazarbayev University — Kazakhstan

Abstract. This paper studies the informativeness of linguistic properties such as part-of-speech and named entities encoded in word representations. First, we find directions that correspond to these properties using the method of Elazar et al. (2020). Then such directions are compared with the principal vectors obtained from application of PCA to word embeddings. As a result, we find that the part-of-speech information is more important for word embeddings than the named entity property.

1 Introduction

Word embeddings became one of the fundamental components of modern natural language processing (NLP) models. To effectively use them in NLP tasks, it is essential to understand what type of linguistic information they learn. Given impressive results on downstream tasks such as machine translation, language modeling, sentiment analysis, etc., the existing studies (e.g., [1, 2]) suggest the presence of semantic, morphological, lexical, syntactic, and other linguistic properties encoded in vector representations of words.

However, Elazar et al. [3] show that there is also redundant information in the representations when predicting a particular property, and it is unclear how these properties are entangled with each other.

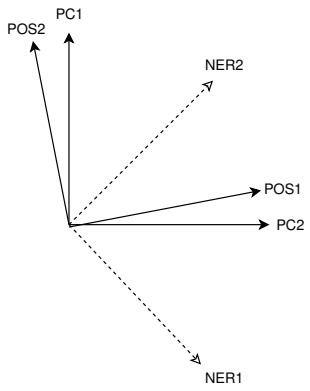

Fig. 1: We show that POS-directions are more aligned with principal component (PC) directions than NER-directions.

To add some clarity to the issue, in this work (following [4] and [3]), we try to find which linguistic property has higher importance for word embeddings. We do so by linearly removing those properties from word vectors and studying their informativeness with Principal Component Analysis (PCA). Similar to [4], our work focuses on linguistic properties such as part-of-speech (POS) and named entity information (NER). We apply PCA on pre-trained word embeddings and find cosine similarity between principal directions and properties directions. We find greater importance of POS directions in the word representations compared to NER directions (Fig. 1), which is consistent with results of [3, 5].

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

2 Related Work

Measuring to what extent linguistic features are present in word embeddings is usually done by probing. [3] proposed *amnesic probing*, which quantifies the importance of linguistic properties (POS, NER, dependency information) by linearly removing the information from word vectors. The difference from this work is that we find the relative impact of linguistic features. [4] conducts experiments in a similar fashion, backing up their empirical findings with a theoretical background and reporting the greater impact of POS information compared to NE. Unlike both of the studies, the benefit of our geometric perspective is that (after removing linguistic information from the embeddings) we do not need to measure the drop in language model performance to examine the significance of linguistic features in word embeddings. There is also an attempt [5] to explore the ability to predict POS tags with correlation analysis with PCA on Czech word embeddings.

3 Background

3.1 Word Representations

Skip-gram with negative sampling (SGNS) architecture, proposed by [6], is trained to predict contextual words given the current word. It is a shallow two-layer neural network, which proved itself powerful by finding semantically and syntactically close words which is evidenced by linear vector operations such as $\mathbf{w}_{\text{king}} - \mathbf{w}_{\text{man}} + \mathbf{w}_{\text{woman}} \approx \mathbf{w}_{\text{queen}}$. We pre-train SGNS on `text8` dataset.[1].
Bidirectional Encoder Representations from Transformers (BERT) [7] is a deep Transformer [8] encoder trained jointly as a masked language model and on next-sentence prediction, trained on the concatenation of the Toronto Books Corpus [9] and English Wikipedia. We use the publicly released pre-trained BERT Base (12-layer) from HuggingFace.

3.2 Geometric Methods

Principal Component Analysis (PCA) is a dimensionality reduction method, which retains most of the variation in the data set. Its first principal direction (first loading vector) is the one along which there is the largest variability of data points. This means that of all possible directions, the first loading vector is the most informative. The second loading vector is the most informative out of all vectors that are orthogonal to the first loading vector, etc. See Fig. 2 for the illustration of PCA for a set of two-dimensional vectors.
Iterative Nullspace Projection (INLP) [10] is a method which linearly removes some property T from the embeddings $\mathbf{x} \in \mathbb{R}^d$. INLP neutralizes the ability to linearly predict T from $\mathbf{x}$. It does so by training a sequence of auxiliary models $\tau_1, \ldots, \tau_k$ that predict T from $\mathbf{x}$, interpreting each one as conveying information on unique directions in the latent space that correspond to T, and

[1] Our implementation is available at `https://github.com/zh3nis/SGNS`

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

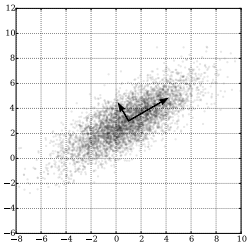

Fig. 2: Principal component analysis.

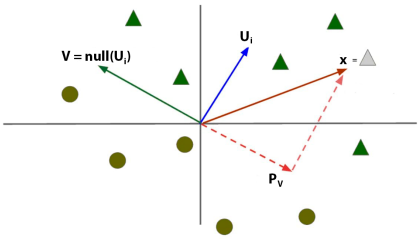

Fig. 3: Iterative nullspace projection.

iteratively removing each of these directions. In the i^{th} iteration, τ_i is a *linear* model[2] parameterized by a matrix $\mathbf{U}_i$ and trained to predict T from $\mathbf{x}$. When the embeddings are projected onto null($\mathbf{U}_i$) by a projection matrix $\mathbf{P}_{\text{null}(\mathbf{U}_i)}$, we have

$$\mathbf{U}_i \mathbf{P}_{\text{null}(\mathbf{U}_i)} \mathbf{x} = \mathbf{0},$$

i.e. τ_i will be unable to predict T from $\mathbf{P}_{\text{null}(\mathbf{U}_i)}\mathbf{x}$ (Fig. 3). Number of iterations k is taken such that no linear classifier achieves above-majority accuracy when predicting T from $\tilde{\mathbf{x}} = \mathbf{P}_{\text{null}(\mathbf{U}_k)} \mathbf{P}_{\text{null}(\mathbf{U}_{k-1})} \dots \mathbf{P}_{\text{null}(\mathbf{U}_1)} \mathbf{x}$.

3.3 Tasks

Part-of-speech tagging (POS) is the task of categorizing each word in a sentence with morpho-syntactic labels. An example of a tagged sentence is given below:

I	want	an	early	upgrade
⟨pronoun⟩	⟨verb⟩	⟨determiner⟩	⟨adjective⟩	⟨noun⟩

Named entity recognition (NER) predicts which words in a text are named entities i.e. persons, locations, and organizations. An example is as follows:

Robert	joined	Microsoft	as	a	data	scientist
⟨person⟩		⟨organization⟩				

Annotated data for both of the above tasks is taken from the English part of the OntoNotes corpus [12]. Intuitively, POS annotation shall contain more information regarding the underlying text than the NER annotation. This is verified in our experiments.

[2] [10] use the Linear SVM [11], and we follow their setup.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

4 Experiments

4.1 Setup

Since the INLP method makes it possible to interpret linguistic properties in terms of directions, and we expect that the POS property is more informative[3] than the NER property, it is natural to hypothesize that the POS directions are *closer* to the principal directions (in cosine similarity) than the NER directions. In what follows, we propose a method for testing this hypothesis.

We perform PCA for the set of 200-dimensional SGNS embeddings, and let $\{\mathbf{v}_i\}_{i=1}^{200}$ be the corresponding principal vectors. Let $\{\mathbf{p}_j\}_{j=1}^{57}$ and $\{\mathbf{n}_j\}_{j=1}^{57}$ be the unit POS and NER vectors as produced by the INLP procedure on the OntoNotes data. The stopping criteria is iterating INLP till the accuracy of a linear classifier drops to the accuracy of a majority-classifier. In case of POS, each of 5 INLP iterations produces 40 directions; and in case of NER, each of 3 INLP iterations produces 19 directions. This gives 200 $\mathbf{p}_j$'s and 57 $\mathbf{n}_j$'s. However, for the sake of fair comparison, we use only the first 57 POS vectors.

Will a set of random vectors be aligned well with the set of principal vectors? To control for this kind of by-chance informativeness we consider a set of random unit vectors $\left\{\frac{\mathbf{r}_j}{\|\mathbf{r}_j\|}\right\}_{j=1}^{57}$, where $\mathbf{r}_j \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$. For constructing confidence bands, we replicate this sampling 100 times.

For the first principal direction $\mathbf{v}_1$ we take the closest (in cosine similarity) vector among $\left\{\mathbf{p}_1, \ldots, \mathbf{p}_{57}, \mathbf{n}_1, \ldots, \mathbf{n}_{57}, \frac{\mathbf{r}_1}{\|\mathbf{r}_j\|}, \ldots, \frac{\mathbf{r}_{57}}{\|\mathbf{r}_{57}\|}\right\}$ and record its type (POS, NER, or Random). We repeat this for the first 57 principal directions, and then report cumulative counts of how many times each of the types was closer to the principal directions.

For BERT, we apply the same procedure with the set of 768-dimensional BERT pre-trained embeddings. In this case, with the same stopping criteria, INLP runs for 7 iterations, which gives us 133 NER vectors. Again, to make the vectors comparable, we take the same number of vectors from all three set of vectors $\left\{\mathbf{p}_1, \ldots, \mathbf{p}_{133}, \mathbf{n}_1, \ldots, \mathbf{n}_{133}, \frac{\mathbf{r}_1}{\|\mathbf{r}_j\|}, \ldots, \frac{\mathbf{r}_{133}}{\|\mathbf{r}_{133}\|}\right\}$.

4.2 Results

The results are provided in Fig. 4. As we can see, in both cases—SGNS and BERT—among POS directions we can find more vectors that are closer to the principal directions than among NER directions, and both are closer aligned with the first 20 principal directions than random vectors.

Bearing in mind that the set of SGNS (BERT) vectors are placed in the Euclidean space in a way that optimizes its objective, we can conclude that the

[3]Here the word *informative* is used in two different ways: a principal vector is informative if there is a large variability of word embeddings along with it; whereas, a POS vector is informative if its removal from word embeddings (by nullspace projection) causes a large increase of a pre-training loss.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

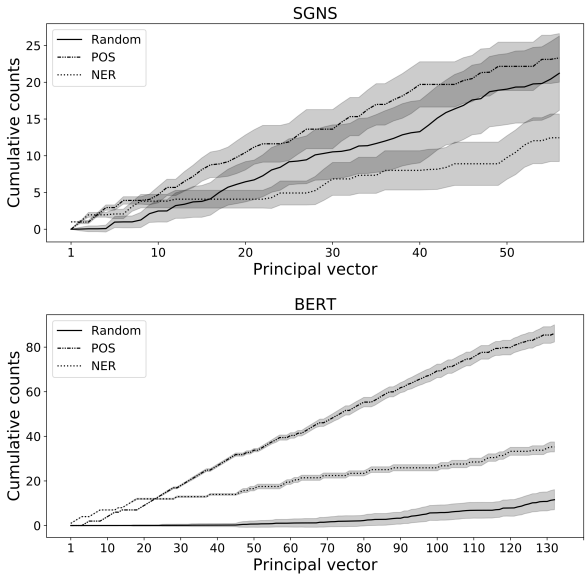

Fig. 4: For each principal direction i we indicate how many times POS-directions, NER-directions, and random directions were closest to the principal directions $[1, 2, \ldots, i]$. Indicated are 90% confidence bands across 100 runs.

POS structure influences stronger such objective than the NER structure, and thus confirms our hypothesis.

5 Conclusion

In this work, we propose a geometric perspective on analyzing the information encoded in word representations. With the help of INLP and PCA algorithms, we were able to make a geometric comparison of two linguistic properties, POS and NER. The importance of POS property proved to be higher than NER for both static and contextual word embeddings. For future work, it would be interesting to examine other debiasing methods on word representations, especially on big models like BERT. There are also other linguistic properties encoded in word vectors which can be further explored by the proposed method.

Acknowledgements

The authors would like to thank Matthias Gallé, Vassilina Nikoulina for the discussion of this work, and the anonymous reviewers for their feedback. This work is supported by the Nazarbayev University Collaborative Research Program 091019CRP2109.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

References

[1] Oded Avraham and Yoav Goldberg. The interplay of semantics and morphology in word embeddings. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of EACL*, pages 422–426, 2017.

[2] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *ICLR, Workshop Track Proceedings*, 2013.

[3] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175, 2021.

[4] Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé, and Zhenisbek Assylbekov. The rediscovery hypothesis: Language models need to meet linguistics. *CoRR*, abs/2103.01819, 2021.

[5] Tomás Musil. Examining structure of word embeddings with PCA. In Kamil Ekstein, editor, *Proceedings of TSD*, volume 11697 of *Lecture Notes in Computer Science*, pages 211–223. Springer, 2019.

[6] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Proceedings of NeurIPS*, pages 3111–3119, 2013.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Proceedings of NeurIPS*, pages 5998–6008, 2017.

[9] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV*, pages 19–27, 2015.

[10] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of ACL*, pages 7237–7256, 2020.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[12] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D Hwang, Claire Bonial, et al. Ontonotes release 5.0. 2012.