Paraphrases do not explain word analogies

Louis Fournier  
Cognitive Machine Learning  
CNRS-EHESS-INRIA  
PSL Research University  
Paris, France  
louis.fournier@polytechnique.edu  

Ewan Dunbar  
Cognitive Machine Learning  
CNRS-EHESS-INRIA  
PSL Research University  
Paris, France  
and  
University of Toronto  
Toronto, Canada  
ewan.dunbar@utoronto.ca

Abstract

Many types of distributional word embeddings (weakly) encode linguistic regularities as directions (the difference between *jump* and *jumped* will be in a similar direction to that of *walk* and *walked*, and so on). Several attempts have been made to explain this fact. We respond to Allen and Hospedales’ recent (ICML, 2019) theoretical explanation, which claims that word2vec and GloVe will encode linguistic regularities whenever a specific relation of paraphrase holds between the four words involved in the regularity. We demonstrate that the explanation does not go through: the paraphrase relations needed under this explanation do not hold empirically.

1 Introduction

The study of linguistic regularities in distributional word embeddings—that the difference vector calculated between the vectors *jump* and *jumped* shows a similar direction to that of *walk* and *walked*, and so on—has been both stimulating and controversial. While a number of such regularities appear to hold, across a number of different kinds of embeddings, the standard 3COSADD analogy test used to measure the presence of these regularities has come under fire for confounding analogical regularities with unrelated properties of semantic embeddings. It is thus important to note that several papers have proposed theoretical explanations for why linguistic regularities *should* hold in distributional word embeddings. Particularly in light of the controversies over linguistic regularities, it is important to examine the soundness of these arguments.

Allen and Hospedales (2019) develop such an explanation by linking the semantic definition of an analogy to paraphrases. In the sense of Gittens et al. (2017), paraphrases are sets of words which are semantically and distributionally closely equivalent to another word or set of words—for example, *king* may be paraphrased by {*man, royal*}. Allen and Hospedales argue that the standard analogy criterion, that *king* - *man* + *woman* = *queen*, is equivalent to a criterion whereby {*king, woman*} paraphrases {*man, queen*}. With this in mind, it becomes possible to rewrite the arithmetic analogy criterion in terms of vectors encoding the pointwise mutual information (PMI) between words and their contexts, and to decompose the error in the analogy equality into several components, including a paraphrase error term measuring the degree to which the critical paraphrase holds. Making use of an assumption that the word2vec embedding is a linear transformation of the PMI matrix, they argue that results in terms of PMI apply to word vectors. Thus, under their explanation, a major part of success on an analogy *a* - *a* + *b* = *b* is due to *a*, *b* and *a*, *b* being close distributional paraphrases.

We first review the literature on the analogy test itself, underlining known pitfalls which any explanation of linguistic regularities must navigate. We then show empirically that the relation between the PMI matrix and word2vec embeddings is to some degree linear, which may be enough to satisfy the assumption of Allen and Hospedales (2019). We further examine the proposed decomposition into error terms. We demonstrate that, empirically, these error terms tend to be undefined due to data sparseness, undermining their explanatory force. Most importantly, examining a number of analogies which pass the standard test, we show that the critical paraphrase error term is, contrary to the proposed explanation, very large.¹

¹Code is available at www.github.com/bootphon/paraphrases_do_not_explain_analogies.
2 Related work

Early works proposing explanations of the analogical properties of word embeddings include Mikolov et al. (2013b) and Pennington et al. (2014). A geometrical explanation is proposed by Arora et al. (2016), but this explanation relies on very strong preconditions, notably, that the word vectors be distributed uniformly in space. Ethayarajh et al. (2019) also propose an explanation, providing a link between the PMI and the norm of word embeddings. However, as pointed out by Allen and Hospedales (2019), this explanation, too, rests on strong assumptions. Notably, the words involved in the analogy are required to be coplanar, a property that seems unlikely in light of the lack of parallelism we discuss in the next section.

3 Issues with the test

Issues have arisen with the standard way of measuring linguistic analogies. Levy and Goldberg (2014a), Vylomova et al. (2016), Rogers et al. (2017), and Fournier et al. (2020) all demonstrate that the standard 3COSADD measure conflates several very different properties of embeddings, simultaneously measuring not only the directional regularities suggested by typical illustrations of vectors in a parallelogram, but also the similarity of individual matched pairs such as king, man, as well as the global arrangement of vectors in semantic fields, such as king, queen, prince, . . . versus man, woman, child, . . . in distinct regions of the space. These issues undermine the construct validity of the standard analogy test. This conflation of properties explains certain pathological behaviours of the test (Linzen, 2016; Rogers et al., 2017). In spite of these issues, Fournier et al. (2020) demonstrate, using alternative measures, that linguistic regularities are nevertheless coded by directional similarities. This parallelism is weak, with directions tending to be closer, in the absolute, to being orthogonal than to being parallel, but is present above chance level (unmatched word pairs).

Thus, before turning to Allen and Hospedales (2019), one of a number of theoretical attempts to explain performance on the 3COSADD objective, we underscore that such demonstrations run the risk of explaining properties of the test which may be of secondary interest, or, conversely, of placing undue emphasis on the role of directional regularities, which have been shown to play only a small role in success on 3COSADD.

4 Explaining analogies through paraphrases

For a word \(w_i\) and a word \(c_j\) which can appear in the context of \(w_i\), the pairwise mutual information \(PMI(w_i, c_j)\) is defined as \(\log \frac{p(w_i, c_j)}{p(w_i)p(c_j)}\). As shown by Levy and Goldberg (2014b), skip-gram word2vec with negative sampling factorizes the PMI: \(PMI \approx W^T \cdot C\), with \(W\) and \(C\) the word and context embedding matrices of a word2vec model.

For two pairs of words \((a, a^*)\) and \((b, b^*)\) from the same semantic relation, the standard arithmetic analogy test criterion is that \(a - a^* + b^* = b\). Writing \(W = \{a, b^*\}\), \(W_s = \{a^*, b\}\), and \(PMI_x\) the PMI vector of \(x\), Allen and Hospedales (2019) show that is possible to rewrite the arithmetic analogy formula with PMI vectors, and to decompose the error in the equality into five terms as follows:

\[
PMI_{b^*} = PMI_b + PMI_{a^*} - PMI_a \\
+ \left(\rho_i W_i W_s + \sigma_i W_s \right) \text{ Paraphrase error} \\
+ \left(\tau_i W - \tau_i W_s \right) \text{ Conditional dependence error} \\
+ \left(\rho_i W_s \right) \text{ Mutual dependence error}
\]

The error terms are vectors of length \(|V|\) (vocabulary size), with each element \(j\) defined as:

\[
\rho_i W_i W_s = \log \frac{p(c_j|W_s)}{p(c_j|W)} \\
\sigma_i W = \log \frac{p(W_i|c_j)}{\prod_{W_i} p(w_i|c_j)} \\
\tau_i W = \log \frac{p(W)}{\prod_{W} p(w_i)}
\]

The authors claim that these terms can be embedded linearly into a word2vec embedding space by multiplying them by the Moore-Penrose pseudoinverse \(C^\dagger\) of the context matrix \(C\). Then with \(w_x\) the word2vec embedding of \(x\), \(C^\dagger \cdot PMI_x \approx w_x\). Thus we get the final decomposition:

\[
w_{b^*} = w_b + w_{a^*} - w_a + C^\dagger \left(\rho_i W_i W_s + \sigma_i W_s - (\tau_i W - \tau_i W_s)\right) 1
\]

The paraphrase error term is claimed to be small for successful analogies. Elaborating on the notation, \(W\) is taken to paraphrase \(W_s\) if, wherever
all $w \in \mathcal{W}$ appear together, we observe the same distribution of surrounding words as for $\mathcal{W}_*$. The paraphrase error assesses the similarity of the distributions of words in the context of $\mathcal{W}$ (all words in $\mathcal{W}$ appearing together) versus $\mathcal{W}_*$.

5  **Linearity of the link between PMI and word2vec**

Though it is true that there is a relation between the word2vec matrices $W^\top \cdot C$ and the PMI matrix, in practice the link is more complicated than simple linear matrix factorization, due in part to the training tricks described in Mikolov et al. (2013a). The result of Allen and Hospedales (2019) requires that the embedding from PMI vectors to word2vec embeddings be “linear enough” for $C^\dagger \cdot \text{PMI}$ to approximate $W$.

To assess this, we use the text8 corpus both to train word2vec embeddings and to estimate a PMI matrix. We replace infinite values in the PMI matrix by 0. In Figure 1a, we show the distribution of the Pearson correlation coefficient (assessing the presence of a linear relation) between the word2vec embedding and the corresponding row of $C^\dagger \cdot \text{PMI}$ for the top ten thousand words in the corpus. As can be seen from the figure, the correlation tends to be between 0.5 and 0.8. For instance in Figure 1b, the word2vec embedding for *king* is plotted against the row of $C^\dagger \cdot \text{PMI}$ corresponding to *king*. While the relation is not perfectly linear—many words have a correlation of around 0.55, far lower than that of *king*—the empirical relations shown here leave open the possibility that it may indeed be “sufficiently linear” to be taken for granted. However, while linearity is necessary for the result of Allen and Hospedales (2019) to go through, it is not sufficient. In the next section, we assess the critical question of whether the paraphrase error is small enough to serve as an explanation for the success of linguistic analogies.

6  **Empirical analysis of the error terms**

We now seek to examine the proposed explanation by calculating the proposed error terms empirically. However, in practice, many of the terms are undefined, since they rely on cooccurrences unattested in practical corpora. The most extreme situation occurs when the two words of a paraphrase $\mathcal{W} = \{w_1, w_2\}$ are never present in the same context window in the corpus. We found that only 16% of the paraphrase sets associated with the BATS analogy set (Gladkova et al., 2016)—for example, *king, woman*—were present together in the text8 corpus in a context window of length five. We refer to such paraphrase sets as “well-defined” with respect to the corpus. The problem of zero co-occurrence counts was anticipated by Allen and Hospedales (2019), who propose to restrict their analysis to the case where the context window is sufficiently large that all relevant terms are well defined. We stress that our trained word2vec vectors are also trained with a context window of five, and yield expected levels of performance on the BATS analogy test, despite having access to little training data on which to model co-occurrences such as *king, woman*, *queen, man*, and so on.

At a minimum, if the proposed explanation holds, the cases for which the error terms are empirically well-defined should show signs of the paraphrase
error being relatively small. We now detail how we implemented the error terms in cases for which they were well-defined. We count co-occurrences \( N(w_i, w_j, w_k) \) in text8 for all triplets of words \( w_i, w_j, w_k \), with \( w_k \) at the center of the context window, and \( W = \{w_i, w_j\} \) any paraphrase, both occurring anywhere within a context window of width five. We restrict analysis to the ten thousand most frequent word types \( w_i \) and \( w_j \), yielding \( 10^8 \) possible paraphrases.\(^4\) We use the relative frequencies as estimators of \( p(w_k|\{w_i, w_j\}) \) and \( p(\{w_i, w_j\}|w_k) \), and marginalize to obtain \( p(w_i|w_k) \), \( p(\{w_i, w_j\}|w_k) \) and \( p(w_k) \). The error terms follow. Since this can still lead to ill-defined elements, we replace \( \log(+\infty) \) and \( \log(0) \) by \( +/ - \log(\epsilon) \), with \( \epsilon = 10^{-15} \) (within reason, the value of \( \epsilon \) is immaterial). We also replace \( \log(0)/0 \) with 0.

Table 1 shows the mean and median values of the L2 norms of the paraphrase error vectors across several categories of the BATS dataset. We compare them with the sum of the four dependence error terms (the dependence error reflects statistical dependencies within \( W \) and \( W_e \), irrelevant to the analogy), as well as the sum of all five error terms (equal to the difference between the PMI of \( W \) and \( W_e \)). The paraphrase error is indeed smaller than the other error terms. However, as we now show, the paraphrase error is not small enough to contribute substantially to the success of analogies.\(^5\)

Take the norm of the paraphrase error vector \( \rho \) as a measure of the divergence in the PMI between two paraphrases. For an analogy with associated paraphrases \( W \) and \( W_e \), we assess how many paraphrases are closer to \( W \) than to \( W_e \) by calculating the rank of the norm of \( \rho^W \cdot W \) among all \( \rho^W \cdot X \), where \( X \) spans over all pairs of words constructible from the top ten thousand most frequent words in the corpus. To do so, we define a Paraphrase Conditional Information matrix (PCI). For \( W_{ij} = \{w_i, w_j\} \) and \( w_k \), we define PCI(i,j, k), the value at column \( li_j \) and row \( k \) to be \( \log(p(W_{ij}|w_k)) \), where with \( l_i \) is a unique index associate with tuple \((i,j)\). We compute only the positive PCI, to obtain a sparse matrix. The difference between two PCI columns is a paraphrase error vector, and their Euclidean distance is the norm of the paraphrase error.

We now compute, for each analogy, the distance between the PCI column of \( W \) and every other column (paraphrase) of the PCI matrix. We calculate the rank of the true analogy pair \( W_e \). Given that the analogy test generally succeeds in picking out \( b \) as being the most similar to \( a - a^* + b^* \) out of the entire vocabulary (modulo Linzen 2016), we would expect that, for successful analogies, the paraphrase error for the true analogy would be among the highest, if small paraphrase error were the explanation for success. Table 2 displays the mean of this rank within each BATS category. The rank is extremely low (in the millions), making the paraphrase error in true analogies far too high to be the explanation for their success.\(^6\)

---

\(^4\)\(w_k\) is allowed to vary over all of the types included in the training for word2vec, of which there are 71290. Thus, for each paraphrase, the error vectors have 71290 elements, one for each vocabulary word.

\(^5\)We note also that the error values seem relatively consistent between categories, while success on the analogy test varies greatly between categories.

\(^6\)Limiting the search to the paraphrases composed by at least one of the words of \( W_e \) still results in a very low rank
7 Conclusion

Recent work has shown that, in spite of the standard analogy test’s confound with simple vector similarity, distributional word vectors genuinely do encode linguistic regularities as directional regularities above and beyond vector similarity (Fournier et al., 2020). Further research is warranted into the mechanisms by which distributional word embeddings come to show these regularities. However, the analysis of analogies as paraphrases does not hold up as an explanation of performance on the analogy test—nor would an explanation of performance on the 3COSADD analogy test be a satisfying result, since the test is not a useful measure to begin with.

Acknowledgments

This work was funded in part by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-17-CE28-0009 GEOMPHON, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute, ANR-18-IDEX-0001 U de Paris, ANR-10-LABX-0083 EFL) and grants from CIFAR (Learning in Machines and Brains), Facebook AI Research (Research Grant), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).

References

Carl Allen and Timothy Hospedales. 2019. Analogies Explained: Towards Understanding Word Embeddings. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 223–231, Long Beach, California, USA. PMLR.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A Latent Variable Model Approach to PMI-based Word Embeddings. Transactions of the Association for Computational Linguistics, 4:385–399.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards Understanding Linear Word Analogies. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. 2020. Analogies minus analogy test: measuring regularities in word embeddings. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 365–375. Online. Association for Computational Linguistics.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. 2017. Skip-gram à’“ Zipf + uniform = vector additivity. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsumo. 2016. Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn’t. In Proceedings of the NAACL-HLT SRW, pages 47–54, San Diego, California, June 12-17, 2016. ACL.

Omer Levy and Yoav Goldberg. 2014a. Linguistic Regularities in Sparse and Explicit Word Representations. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 2177–2185. Curran Associates, Inc.

Tal Linzen. 2016. Issues in Evaluating Semantic Spaces Using Word Analogies. In Proceedings of the First Workshop on Evaluating Vector Space Representations for NLP. Association for Computational Linguistics.

Matt Mahoney. 2006. Large text compression benchmark: About the test data (web page).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In Proceedings of International Conference on Learning Representations (ICLR).

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL-HLT 2013, pages 746–751, Atlanta, Georgia, 9–14 June 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. 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. Association for Computational Linguistics.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (Too Many) Problems of Analogical Reasoning with Word Vectors. In Proceedings of the 6th Joint
Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1671–1682, Berlin, Germany. Association for Computational Linguistics.