Towards Understanding Linear Word Analogies

Kawin Ethayarajh, David Duvenaud[†], Graeme Hirst
University of Toronto
[†]Vector Institute
{kawin, duvenaud, gh}@cs.toronto.edu

Abstract

A surprising property of word vectors is that word analogies can often be solved with vector arithmetic. However, it is unclear why arithmetic operators correspond to non-linear embedding models such as skip-gram with negative sampling (SGNS). We provide a formal explanation of this phenomenon without making the strong assumptions that past theories have made about the vector space and word distribution. Our theory has several implications. Past work has conjectured that linear substructures exist in vector spaces because relations can be represented as ratios; we prove that this holds for SGNS. We provide novel justification for the addition of SGNS word vectors by showing that it automatically downweights the more frequent word, as weighting schemes do *ad hoc*. Lastly, we offer an information theoretic interpretation of Euclidean distance in vector spaces, justifying its use in capturing word dissimilarity.

1 Introduction

Distributed representations of words are a cornerstone of current methods in natural language processing. Word embeddings, also known as word vectors, can be generated by a variety of models, all of which share Firth's philosophy (1957) that the meaning of a word is defined by "the company it keeps". The simplest such models obtain word vectors by constructing a low-rank approximation of a matrix containing a co-occurrence statistic (Landauer and Dumais, 1997; Rohde et al., 2006). In contrast, neural network models (Bengio et al., 2003; Mikolov et al., 2013b) learn word embeddings by trying to predict words using the contexts they appear in, or vice-versa.

A surprising property of word vectors learned via neural networks is that word analogies can often be solved with vector arithmetic. For example,

'king is to ? as man is to woman' can be solved by finding the closest vector to $\vec{king} - \vec{man} + \vec{woman}$, which should be *queen*. It is unclear why arithmetic operators can effectively compose embeddings generated by non-linear models such as skip-gram with negative sampling (SGNS). There have been two attempts to rigorously explain this phenomenon, but both have made strong assumptions about either the embedding space or the word distribution. The paraphrase model (Gittens et al., 2017) hinges on words having a uniform distribution rather than the typical Zipf distribution, which the authors themselves acknowledge is unrealistic. The latent variable model (Arora et al., 2016) assumes that word vectors are known *a priori* and generated by randomly scaling vectors sampled from the unit sphere.

In this paper, we explain why – and under what conditions – word analogies can be solved with vector arithmetic, without making the strong assumptions past work has. We focus on GloVe and SGNS because they implicitly factorize a word-context matrix containing a co-occurrence statistic (Levy and Goldberg, 2014), which allows us to interpret the inner product of a word and context vector. We begin by formalizing word analogies as functions that transform one word vector into another. When this transformation is simply the addition of a displacement vector – as is the case when using vector arithmetic – we call the analogy a *linear analogy*. Central to our theory is the expression $PMI(x, y) + \log p(x, y)$, which we call the *co-occurrence shifted pointwise mutual information* (csPMI) of (x, y).

We prove that in both SGNS and GloVe spaces without reconstruction error (i.e., when the factorized word-context matrix can be perfectly reconstructed), a linear analogy holds over a set of ordered word pairs iff $csPMI(x, y)$ is the same for every word pair, $csPMI(x_1, x_2) = csPMI(y_1, y_2)$

for any two word pairs, and the row vectors of x_1, x_2, y_1, y_2 in the factorized matrix are coplanar. By then framing vector addition as a kind of word analogy, we offer several new insights:

1. Past work has often cited the Pennington et al. (2014) conjecture as an intuitive explanation of why vector arithmetic works for analogy solving. The conjecture is that an analogy of the form *a is to b as x is to y* holds iff $p(w|a)/p(w|b) \approx p(w|x)/p(w|y)$ for every word w in the vocabulary. While this is sensible, it is not based on any theoretical derivation or empirical support. We provide a formal proof that this is indeed true.

2. Consider two words x, y and their sum $\vec{z} = \vec{x} + \vec{y}$ in an SGNS embedding space with no reconstruction error. If z were in the vocabulary, the similarity between z and x (as measured by the csPMI) would be the log probability of y shifted by a model-specific constant. This implies that the addition of two words automatically down-weights the more frequent word. Since many weighting schemes are based on the idea that more frequent words should be down-weighted *ad hoc* (Arora et al., 2017), the fact that this is done automatically provides novel justification for using addition to compose words.

3. Consider any two words x, y in an SGNS or GloVe embedding space with no reconstruction error. The squared Euclidean distance between $\vec{x}$ and $\vec{y}$ is a decreasing linear function of $csPMI(x, y)$. In other words, the more similar two words are (as measured by csPMI) the smaller the distance between their vectors. Although this is intuitive, it is also the first rigorous explanation of why the Euclidean distance in embedding space is a good proxy for word dissimilarity.

Although our main theorem only concerns embedding spaces with no reconstruction error, we also explain why, in practice, linear word analogies hold in embedding spaces with some noise. We conduct experiments that support the few assumptions we make and show that the transformations represented by various word analogies correspond to different csPMI values. Without making the strong assumptions of past theories, we thus offer a formal explanation of why, and when, word analogies can be solved with vector arithmetic.

2 Related Work

PMI Pointwise mutual information (PMI) captures how much more frequently x, y co-occur than by chance (Church and Hanks, 1990):

$$\text{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Word Embeddings Word embeddings are distributed representations in a low-dimensional continuous space. Also called word vectors, they capture semantic and syntactic properties of words, even allowing relationships to be expressed arithmetically (Mikolov et al., 2013b). Word vectors are generally obtained in two ways: (a) from neural networks that learn representations by predicting co-occurrence patterns in the training corpus (Bengio et al., 2003; Mikolov et al., 2013b; Collobert and Weston, 2008); (b) from low-rank approximations of word-context matrices containing a co-occurrence statistic (Landauer and Dumais, 1997; Levy and Goldberg, 2014).

SGNS The objective of skip-gram with negative sampling (SGNS) is to maximize the probability of observed word-context pairs and to minimize the probability of k randomly sampled negative examples. For an observed word-context pair (w, c), the objective would be $\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c' \sim P_n}[\log(-\vec{w} \cdot \vec{c'})]$, where c' is the negative context, randomly sampled from a scaled distribution P_n. Though no co-occurrence statistics are explicitly calculated, Levy and Goldberg (2014) proved that SGNS is in fact implicitly factorizing a word-context PMI matrix shifted by $-\log k$.

Latent Variable Model The latent variable model (Arora et al., 2016) was the first attempt at rigorously explaining why word analogies can be solved arithmetically. It is a generative model that assumes that word vectors are generated by the random walk of a "discourse" vector on the unit sphere. Gittens et al.'s criticism of this proof is that it assumes that word vectors are known *a priori* and generated by randomly scaling vectors uniformly sampled from the unit sphere (or having properties consistent with this sampling procedure). The theory also relies on word vectors being uniformly distributed (isotropic) in embedding space; however, experiments by Mimno and Thompson (2017) have found that this generally does not hold in practice, at least for SGNS.

Paraphrase Model The paraphrase model (Gittens et al., 2017) was the only other attempt to formally explain why word analogies can be solved arithmetically. It proposes that any set of context words $C = \{c_1, ..., c_m\}$ is semantically equivalent to a single word c if $p(w|c_1, ..., c_m) = p(w|c)$. One problem with this is that the number of possible context sets far exceeds the vocabulary size, precluding a one-to-one mapping; the authors circumvent this problem by replacing exact equality with the minimization of KL divergence. Assuming that the words have a uniform distribution, the paraphrase of C can then be written as an unweighted sum of its context vectors. However, this uniformity assumption is unrealistic – word frequencies obey a Zipf distribution, which is Pareto (Piantadosi, 2014). A later attempt at using paraphrases (Allen and Hospedales, 2019) completely ignores the effect of negative sampling in SGNS' factorization. Neither work provides any empirical evidence in support of the paraphrase model.

3 The Structure of Word Analogies

3.1 Formalizing Analogies

A word analogy is a statement of the form *"a is to b as x is to y"*, which we will write as *(a,b)::(x,y)*. It asserts that a and x can be transformed in the same way to get b and y respectively, and that b and y can be inversely transformed to get a and x. A word analogy can hold over an arbitrary number of ordered pairs: e.g., *"Berlin is to Germany as Paris is to France as Ottawa is to Canada ..."*. The elements in each pair are not necessarily in the same space – for example, the transformation for *(king,roi)::(queen,reine)* is English-to-French translation. For *(king,queen)::(man,woman)*, the canonical analogy in the literature, the transformation corresponds to changing the gender. Therefore, to formalize the definition of an analogy, we will refer to it as a transformation.

Definition 1 *An analogy f is an invertible transformation that holds over a set of ordered pairs S iff $\forall (x,y) \in S, f(x) = y \wedge f^{-1}(y) = x$.*

The word embedding literature (Mikolov et al., 2013b; Pennington et al., 2014) has focused on a very specific type of transformation, the addition of a displacement vector. For example, for *(king,queen)::(man,woman)*, the transformation would be $\vec{king} + (\vec{woman} - \vec{man}) = \vec{queen}$, where the displacement vector is expressed as the difference $(\vec{woman} - \vec{man})$. To make a distinction with our general class of analogies in Definition 1, we will refer to these as *linear analogies*.

Definition 2 *A linear analogy f is an invertible transformation of the form $\vec{x} \mapsto \vec{x} + \vec{r}$. f holds over a set of ordered pairs S iff $\forall (x,y) \in S, \vec{x} + \vec{r} = \vec{y}$.*

Definition 3 *Let W be an SGNS or GloVe word embedding space and C its corresponding context space. Let k denote the number of negative samples, $X_{x,y}$ the frequency, and b_x, b_y the learned biases for GloVe. If there is no reconstruction error, for any words x, y with $\vec{x}, \vec{y} \in W$ and $\vec{x_c}, \vec{y_c} \in C$:*

$$
\begin{aligned}
\text{SGNS}: \quad & \langle \vec{x}, \vec{y_c} \rangle = \text{PMI}(x,y) - \log k \\
\text{GloVe}: \quad & \langle \vec{x}, \vec{y_c} \rangle = \log X_{x,y} - b_x - b_y
\end{aligned}
\tag{2}
$$

SGNS and GloVe generate two vectors for each word in the vocabulary: a context vector, for when it is a context word, and a word vector, for when it is a target word. Context vectors are generally discarded after training. The SGNS identity in (2) is from Levy and Goldberg (2014), who proved that SGNS is implicitly factorizing the shifted word-context PMI matrix. The GloVe identity is simply the local objective for a word pair (Pennington et al., 2014). Since the matrix being factorized in both models is symmetric, $\langle \vec{x}, \vec{y_c} \rangle = \langle \vec{x_c}, \vec{y} \rangle$.

Definition 4 *The co-occurrence shifted PMI of a word pair (x,y) is $PMI(x,y) + \log p(x,y)$.*

Definition 5 *Let M denote the word-context matrix that is implicitly factorized by GloVe or SGNS. If there is no reconstruction error, any four words $\{a,b,x,y\}$ are contextually coplanar iff*

$$
\text{rank}\left(\begin{bmatrix} M_{a,\cdot} - M_{y,\cdot} \\ M_{b,\cdot} - M_{y,\cdot} \\ M_{x,\cdot} - M_{y,\cdot} \end{bmatrix} \right) \leq 2
\tag{3}
$$

For example, for SGNS, the first row of this matrix would be $(\text{PMI}(a,\cdot) - \log k) - (\text{PMI}(y,\cdot) - \log k) = \log[p(\cdot|a)/p(\cdot|y)]$. This condition can be trivially derived from the fact that any four vectors $\vec{a}, \vec{b}, \vec{x}, \vec{y}$ in a d-dimensional space (for $d \geq 3$) are coplanar iff $\text{rank}(W^*) \leq 2$, where

$$
W^* = \begin{bmatrix} \vec{a}^T - \vec{y}^T \\ \vec{b}^T - \vec{y}^T \\ \vec{x}^T - \vec{y}^T \end{bmatrix}
\tag{4}
$$

Given that the vocabulary size is much greater than the dimensionality d, and assuming that the

context matrix C is full rank, $\text{rank}(W^*C^T) = \text{rank}(W^*)$. The product W^*C^T is the matrix in (3); each of its three rows is the difference between two rows of M (e.g., $M_{a,\cdot} - M_{y,\cdot}$). Thus we can translate coplanarity in the embedding space to the coplanarity of M's row vectors.

Co-occurrence Shifted PMI Theorem *Let W be an SGNS or GloVe word embedding space with no reconstruction error and S be a set of ordered word pairs such that $\forall\ (x,y) \in S, \vec{x}, \vec{y} \in W$ and $|S| > 1$. A linear analogy f holds over S iff $\exists\ \gamma \in \mathbb{R}$, $\forall (x,y) \in S, csPMI(x,y) = \gamma$ and for any two word pairs $(x_1, y_1), (x_2, y_2) \in S$, the four words are contextually coplanar and $csPMI(x_1, x_2) = csPMI(y_1, y_2)$.*

In sections 3.2 to 3.4 of this paper, we prove the csPMI Theorem. In section 3.5, we explain why, in practice, perfect reconstruction is not needed to solve word analogies using vector arithmetic. In section 4, we explore what the csPMI Theorem implies about vector addition and Euclidean distance in embedding spaces.

3.2 Analogies as Parallelograms

Lemma 1 *A linear analogy f holds over a set of ordered word pairs S iff $\exists\ \gamma' \in \mathbb{R}, \forall\ (x,y) \in S, 2\langle\vec{x},\vec{y}\rangle - \|\vec{x}\|_2^2 - \|\vec{y}\|_2^2 = \gamma'$ and for any two pairs $(x_1, y_1), (x_2, y_2) \in S$, words x_1, x_2, y_1, y_2 are coplanar and $2\langle\vec{x_1}, \vec{x_2}\rangle - \|\vec{x_1}\|_2^2 - \|\vec{x_2}\|_2^2 = 2\langle\vec{y_1}, \vec{y_2}\rangle - \|\vec{y_1}\|_2^2 - \|\vec{y_2}\|_2^2$.*

f holds over every subset $\{(x_1, y_1), (x_2, y_2)\} \subset S$ iff it holds over S. We start by noting that by Definition 2, f holds over $\{(x_1, y_1), (x_2, y_2)\}$ iff:

$$\vec{x_1} + \vec{r} = \vec{y_1} \wedge \vec{x_2} + \vec{r} = \vec{y_2} \qquad (5)$$

By rearranging (5), we know that $\vec{x_2} - \vec{y_2} = \vec{x_1} - \vec{y_1}$ and $\vec{x_2} - \vec{x_1} = \vec{y_2} - \vec{y_1}$. Put another way, x_1, y_1, x_2, y_2 form a quadrilateral in vector space whose opposite sides are parallel and equal in length. By definition, this quadrilateral is then a parallelogram. In fact, this is often how word analogies are visualized in the literature (see Figure 1).

To prove the first part of Lemma 1, we let $\gamma' = -\|\vec{r}\|_2^2$. A quadrilateral is a parallelogram iff each pair of opposite sides is equal in length. For every possible subset, $\vec{r} = (\vec{y_1} - \vec{x_1}) = (\vec{y_2} - \vec{x_2})$. This implies that $\forall\ (x,y) \in S$,

$$\gamma' = -\|\vec{y} - \vec{x}\|_2^2 = 2\langle\vec{x}, \vec{y}\rangle - \|\vec{x}\|_2^2 - \|\vec{y}\|_2^2 \qquad (6)$$

However, this condition is only necessary and not sufficient for the parallelogram to hold. The other
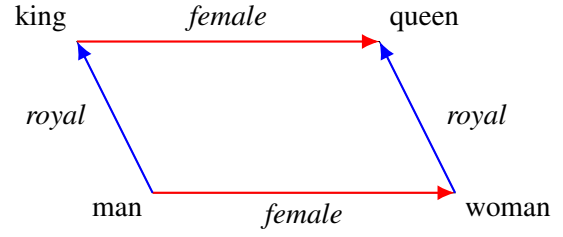

Figure 1: The parallelogram structure of the linear analogy *(king,queen)::(man,woman)*. A linear analogy transforms the first element in an ordered word pair by adding a displacement vector to it. Arrows indicate the directions of the semantic relations.

pair of opposite sides, which do not correspond to $\vec{r}$, are equal in length iff $-\|\vec{x_1} - \vec{x_2}\|_2^2 = -\|\vec{y_1} - \vec{y_2}\|_2^2 \iff 2\langle\vec{x_1}, \vec{x_2}\rangle - \|\vec{x_1}\|_2^2 - \|\vec{x_2}\|_2^2 = 2\langle\vec{y_1}, \vec{y_2}\rangle - \|\vec{y_1}\|_2^2 - \|\vec{y_2}\|_2^2$, as stated in Lemma 1. Note that the sides that do not equal $\vec{r}$ do not necessarily have a fixed length across different subsets of S.

Although points defining a parallelogram are necessarily coplanar, in higher dimensional embedding spaces, it is possible for $\|\vec{x_1} - \vec{x_2}\| = \|\vec{y_1} - \vec{y_2}\|$ and $\|\vec{y_1} - \vec{x_1}\| = \|\vec{y_2} - \vec{x_2}\|$ to be satisfied without the points necessarily defining a parallelogram. Therefore, we must also require that x_1, y_1, x_2, y_2 be coplanar. However, we do not need the word embeddings themselves to verify coplanarity; when there is no reconstruction error, we can express it as a constraint over M, the matrix that is implicitly factorized by the embedding model (see Definition 5).

3.3 Analogies in the Context Space

Lemma 2 *A linear analogy $f : \vec{x} \mapsto \vec{x} + \vec{r}$ holds over a set of ordered pairs S in an SGNS or GloVe word embedding space W with no reconstruction error iff $\exists\ \lambda \in \mathbb{R}, g : \vec{x_c} \mapsto \vec{x_c} + \lambda\vec{r}$ holds over S in the corresponding context space C.*

In other words, an analogy f that holds over S in the word space has a corresponding analogy g that holds over S in the context space. The displacement vector of g is simply the displacement vector of f scaled by some $\lambda \in \mathbb{R}$. To prove this, we begin with (5) and any word w in the vocabulary:

$$
\begin{aligned}
&\vec{x_2} - \vec{y_2} = \vec{x_1} - \vec{y_1} \\
\iff &\langle\vec{w_c}, (\vec{x_2} - \vec{y_2}) - (\vec{x_1} - \vec{y_1})\rangle = 0 \\
\iff &\langle\vec{w}, (\vec{x_{2c}} - \vec{y_{2c}}) - (\vec{x_{1c}} - \vec{y_{1c}})\rangle = 0 \\
\iff &\vec{x_{2c}} - \vec{y_{2c}} = \vec{x_{1c}} - \vec{y_{1c}}
\end{aligned}
\qquad (7)
$$

Note that we can rewrite the second equation as the third because the matrices being factorized in

(2) are symmetric and there is no reconstruction error. We can simplify from the second-last step because not all word vectors lie in the same hyperplane, implying that $(\vec{x}_{2_c} - \vec{y}_{2_c}) - (\vec{x}_{1_c} - \vec{y}_{1_c}) = \vec{0}$.

Thus a linear analogy with displacement vector $(\vec{y}_1 - \vec{x}_1)$ holds over S in the word embedding space iff an analogy with displacement vector $(\vec{y}_{1_c} - \vec{x}_{1_c})$ holds over S in the context space. This is supported by empirical findings that word and context spaces perform equally well on word analogy tasks (Pennington et al., 2014). Since there is an analogous parallelogram structure formed by x_1, y_1, x_2, y_2 in the context space, there is some linear map from $\vec{w} \mapsto \vec{w}_c$ for each word $w \in S$. The real matrix A describing this linear map is symmetric: $\langle \vec{x}, \vec{y}_c \rangle = \vec{x}^T A \vec{y} = (A^T \vec{x})^T \vec{y} = \langle \vec{x}_c, \vec{y} \rangle$ for any $(x, y) \in S$. This implies that $C = AW$, since $\langle \vec{w}, \vec{x}_c \rangle = \langle \vec{w}_c, \vec{x} \rangle$ for any word w.

Since A is a real symmetric matrix, by the finite-dimensional spectral theorem, there is an orthonormal basis of W consisting of eigenvectors of A. If A had distinct eigenvalues, then the relative geometry of the word embeddings would not be preserved by the transformation, in which case it would be possible for two words x, y to satisfy $\langle \vec{x}, \vec{y}_c \rangle \neq \langle \vec{x}_c, \vec{y} \rangle$. This would be a contradiction, given that the factorized word-context matrix is symmetric. Therefore, the relative geometry is only preserved when A has non-distinct eigenvalues. Because A's eigenvectors are a basis for W and all have the same eigenvalue λ, all word vectors lie in the same eigenspace: $\exists \lambda \in \mathbb{R}, \forall \vec{w} \in W, \vec{w}_c = A\vec{w} = \lambda\vec{w}$. Experiments on embedding isotropy in past work (Mimno and Thompson, 2017) provide some empirical support of this result.

3.4 Proof of the csPMI Theorem

From Lemma 1, we know that if a linear analogy f holds over a set of ordered pairs S, then $\exists \gamma' \in \mathbb{R}, \forall (x, y) \in S, 2\langle \vec{x}, \vec{y} \rangle - \|\vec{x}\|_2^2 - \|\vec{y}\|_2^2 = \gamma'$. Because there is no reconstruction error, by Lemma 2, we can rewrite the inner product of two word vectors in terms of the inner product of a word and context vector. Using the SGNS identity in (2), we can then rewrite (6):

$$\begin{aligned}
\gamma' &= 2\langle \vec{x}, \vec{y} \rangle - \|\vec{x}\|_2^2 - \|\vec{y}\|_2^2 \\
&= (1/\lambda)\langle \vec{x} - \vec{y}, \vec{y}_c - \vec{x}_c \rangle \\
\lambda\gamma' &= 2\,\text{PMI}(x, y) - \text{PMI}(x, x) - \text{PMI}(y, y) \\
&= \text{csPMI}(x, y) - \log p(x|x)p(y|y)
\end{aligned} \quad (8)$$

We get the same equation using the GloVe identity in (2), since the learned bias terms b_x, b_y cancel out. Note that $p(x|x) \neq 1$ because $p(x|x)$ is the probability that the word x will appear in the context window when the target word is also x, which is not guaranteed.

For $\log p(x|x)p(y|y)$ to not be undefined, every word in S must appear in its own context at least once in the training corpus. However, depending on the size of the corpus and the context window, this may not necessarily occur. For this reason, we assume that $p(w, w)$, the probability that a word co-occurs with itself, follows the Zipf distribution of $p(w)$ scaled by some constant $\rho \in (0, 1)$. We find this assumption to be justified, since the Pearson correlation between $p(w)$ and non-zero $p(w, w)$ is 0.825 for uniformly randomly sampled words in Wikipedia. We can therefore treat $\log p(x|x)p(y|y) \,\forall\, (x, y) \in S$ as a constant $\alpha \in \mathbb{R}^-$. Rewriting (8), we get

$$\lambda\gamma' + \alpha = \text{csPMI}(x, y) \quad (9)$$

The second identity in Lemma 1 can be expanded analogously, implying that f holds over a set of ordered pairs S iff (9) holds for every pair $(x, y) \in S$ and $\text{csPMI}(x_1, x_2) = \text{csPMI}(y_1, y_2)$ for any two pairs $(x_1, y_1), (x_2, y_2) \in S$ with contextually coplanar words. In section 5, we provide empirical support of this finding by showing that there is a moderately strong correlation (Pearson's $r > 0.50$) between $\text{csPMI}(x, y)$ and γ', in both normalized and unnormalized SGNS embedding spaces.

3.5 Robustness to Noise

In practice, linear word analogies hold in embedding spaces even when there is non-zero reconstruction error. There are three reasons for this: the definition of vector equality is looser in practice, the number of word pairs in an analogy set is small relative to vocabulary size, and analogies mostly hold over frequent word pairs, which are associated with less variance in reconstruction error. For one, in practice, an analogy task $(a, ?)::(x, y)$ is solved by finding the *most similar* word vector to $\vec{a} + (\vec{y} - \vec{x})$, where dissimilarity is defined in terms of Euclidean or cosine distance and $\vec{a}, \vec{x}, \vec{y}$ are excluded as possible answers (Mikolov et al., 2013b). The correct solution to a word analogy can be found even when that solution is not exact. This also means that the solution does not need to lie exactly on the plane defined

by $\vec{a},\vec{x},\vec{y}$. Although the csPMI Theorem assumes no reconstruction error for all word pairs, if we ignore the coplanarity constraint in Definition 5, only $|S|^2+2|S|$ word pairs need to have no reconstruction error for f to hold exactly over S. This number is far smaller than the size of the factorized word-context matrix.

Lastly, in practice, linear word analogies mostly hold over frequent word pairs, which are associated with less variance in reconstruction error. More specifically, for a word pair (x,y), the variance of the noise $\varepsilon_{x,y} = M_{x,y} - \langle \vec{x}, \vec{y_c} \rangle$ is a strictly decreasing function of its frequency $X_{x,y}$. This is because the cost of deviating from the optimal value is higher for more frequent word pairs: this is implicit in the SGNS objective (Levy and Goldberg, 2014) and explicit in GloVe objective (Pennington et al., 2014). We also show that this holds empirically in section 5. Assuming $\varepsilon_{x,y} \sim \mathcal{N}(0,h(X_{x,y}))$, where δ is the Dirac delta distribution:

$$\lim_{X_{x,y} \to \infty} h(X_{x,y}) = 0 \implies \lim_{X_{x,y} \to \infty} \mathcal{N}(0,h(X_{x,y})) = \delta$$
$$\implies \lim_{X_{x,y} \to \infty} \varepsilon_{x,y} = 0$$
$$(10)$$

As the frequency increases, the probability that the noise is close to zero increases. Although word pairs do not have an infinitely large frequency, as long as the frequency of each word pair is sufficiently large, the noise will likely be small enough for a linear analogy to hold over them *in practice*. Our experiments in section 5 bear this out: analogies involving countries and their capitals, which have a median word pair frequency of 3436.5 in Wikipedia, can be solved with 95.4% accuracy; analogies involving countries and their currency, which have a median frequency of just 19, can only be solved with 9.2% accuracy.

A possible benefit of h mapping lower frequencies to larger variances is that it reduces the probability that a linear analogy f will hold over rare word pairs. One way of interpreting this is that h essentially filters out the word pairs for which there is insufficient evidence, even if the conditions in the csPMI Theorem are satisfied. This would explain why reducing the dimensionality of word vectors – up to a point – actually improves performance on word analogy tasks (Yin and Shen, 2018). Representations with the optimal dimensionality have enough noise to preclude spurious analogies that satisfy the csPMI Theorem,

but not so much noise that non-spurious analogies (e.g., *(king,queen)::(man,woman)*) are also precluded.

4   Vector Addition as a Word Analogy

4.1   Formalizing Addition

Corollary 1   *Let $\vec{z} = \vec{x} + \vec{y}$ be the sum of words x,y in an SGNS word embedding space with no reconstruction error. If z were a word in the vocabulary, where δ is a model-specific constant, $csPMI(x,z) = \log p(y) + \delta$.*

To frame the addition of two words x,y as an analogy, we need to define a set of ordered pairs S such that a linear analogy holds over S iff $\vec{x} + \vec{y} = \vec{z}$. To this end, consider the set $\{(x,z),(\emptyset,y)\}$, where z is a placeholder for the composition of x and y and the null word $\emptyset$ maps to $\vec{0}$ for a given embedding space. From Definition 2:

$$(\vec{x} + \vec{r} = \vec{z}) \wedge (\vec{\emptyset} + \vec{r} = \vec{y})$$
$$\iff \vec{z} - \vec{x} = \vec{y} - \vec{\emptyset} \qquad (11)$$
$$\iff \vec{x} + \vec{y} = \vec{z}$$

Even though $\emptyset$ is not in the vocabulary, we can map it to $\vec{0}$ because its presence does not affect any other word vector. To understand why, consider the shifted word-context PMI matrix M that does not have $\emptyset$, and the matrix M' that does, of which M is a submatrix. Where W and C are the word and context matrices, $WC^T = M \iff [W\ \vec{0}][C\ \vec{0}]^T = M'$. Even if the null word does not exist for a given corpus, the embeddings we would get by training on a corpus that did have the null word would otherwise be identical.

An inner product with the zero vector is always 0, so we can infer from the SGNS identity in (2) that $PMI(\emptyset, \cdot) - \log k = 0$ for every word in the vocabulary. The vectors $\vec{x}, \vec{y}, \vec{z}, \vec{\emptyset}$ are all coplanar, and we know from the csPMI Theorem that if a linear analogy holds over $\{(x,z),(\emptyset,y)\}$, then

$$PMI(x,z) + \log p(x,z)$$
$$= 2\,PMI(\emptyset,y) + \log p(y) + \log p(\emptyset)$$
$$= \log p(y) + \delta \qquad (12)$$
$$\text{where } \delta = \log k^2 + \log p(\emptyset)$$

Thus the csPMI of the sum and one word is equal to the log probability of the other word shifted by a model-specific constant. If we assume, as in section 3.5, that the noise is normally distributed, then

3258

even without the assumption of zero reconstruction error, the csPMI of the sum and one word is *on average* equal to the log probability of the other word shifted by a constant. We cannot repeat this derivation with GloVe because it is unclear what the optimal values of the learned biases would be, even with perfect reconstruction.

4.2 Automatically Weighting Words

Corollary 2 *In an SGNS word embedding space, on average, the sum of two words has more in common with the rarer word, where commonality is measured by csPMI.*

For two words x, y, assume without loss of generality that $p(x) > p(y)$. By (12):

$$p(x) > p(y) \iff \log p(x) + \delta > \log p(y) + \delta$$
$$\iff \text{csPMI}(z, y) > \text{csPMI}(z, x) \tag{13}$$

Therefore addition automatically down-weights the more frequent word. For example, if the vectors for $x = $ *'the'* and $y = $ *'apple'* were added to create a vector for $z = $ *'the apple'*, we would expect csPMI(*'the apple'*, *'apple'*) > csPMI(*'the apple'*, *'the'*); being a stopword, *'the'* would on average be heavily down-weighted. While the rarer word is not always the more informative one, weighting schemes like inverse document frequency (IDF) (Robertson, 2004) and unsupervised smoothed inverse frequency (uSIF) (Ethayarajh, 2018) are all based on the principle that more frequent words should be down-weighted because they are typically less informative. The fact that addition automatically down-weights the more frequent word thus provides novel justification for using addition to compose words.

4.3 Interpreting Euclidean Distance

Corollary 3 $\exists \lambda \in \mathbb{R}^+, \alpha \in \mathbb{R}^-$ *such that for any two words x and y in an SGNS or GloVe embedding space with no reconstruction error,* $\lambda \|\vec{x} - \vec{y}\|_2^2 = -\text{csPMI}(x, y) + \alpha$.

From (9), we know that for some $\lambda, \alpha, \gamma' \in \mathbb{R}$, $\text{csPMI}(x, y) = \lambda \gamma' + \alpha$, where $\gamma' = -\|\vec{x} - \vec{y}\|_2^2$. Rearranging this identity, we get

$$\|\vec{x} - \vec{y}\|_2^2 = -\gamma'$$
$$= (-1/\lambda)(\text{csPMI}(x, y) - \alpha) \tag{14}$$
$$\lambda \|\vec{x} - \vec{y}\|_2^2 = -\text{csPMI}(x, y) + \alpha$$

Thus the squared Euclidean distance between two word vectors is simply a linear function of the

negative csPMI. Since $\text{csPMI}(x, y) \in (-\infty, 0]$ and $\|\vec{x} - \vec{y}\|_2^2$ is non-negative, λ is positive. This identity is intuitive: the more similar two words are (as measured by csPMI), the smaller the distance between their word embeddings. In section 5, we provide empirical evidence of this, showing that there is a moderately strong positive correlation (Pearson's $r > 0.50$) between $-\text{csPMI}(x, y)$ and $\|\vec{x} - \vec{y}\|_2^2$, in both normalized and unnormalized SGNS embedding spaces.

4.4 Are Relations Ratios?

Pennington et al. (2014) conjectured that linear relationships in the embedding space – which we call displacements – correspond to ratios of the form $p(w|x)/p(w|y)$, where (x, y) is a pair of words such that $\vec{y} - \vec{x}$ is the displacement and w is some word in the vocabulary. This claim has since been repeated in other work (Arora et al., 2016). For example, according to this conjecture, the analogy *(king,queen)::(man,woman)* holds iff for every word w in the vocabulary

$$\frac{p(w|king)}{p(w|queen)} \approx \frac{p(w|man)}{p(w|woman)} \tag{15}$$

However, as noted earlier, this idea was neither derived from empirical results nor rigorous theory, and there has been no work to suggest that it would hold for models other than GloVe, which was designed around it. We now prove this conjecture for SGNS using the csPMI Theorem.

Pennington et al. Conjecture Let S be a set of ordered word pairs (x, y) with vectors in an embedding space. A linear word analogy holds over S iff $\forall (x_1, y_1), (x_2, y_2) \in S, p(w|x_1)/p(w|y_1) \approx p(w|x_2)/p(w|y_2)$ for every word w in the vocabulary.

Assuming there is no reconstruction error, we replace approximate equality with exact equality and rewrite the identity for SGNS using (2):

$$\frac{p(w|x_1)}{p(w|y_1)} = \frac{p(w|x_2)}{p(w|y_2)}$$
$$\iff \text{PMI}(w, x_1) - \text{PMI}(w, y_1) =$$
$$\text{PMI}(w, x_2) - \text{PMI}(w, y_2)$$
$$\iff \langle \vec{w_c}, \vec{x_1} \rangle - \langle \vec{w_c}, \vec{y_1} \rangle = \langle \vec{w_c}, \vec{x_2} \rangle - \langle \vec{w_c}, \vec{y_2} \rangle$$
$$\iff \langle \vec{w_c}, (\vec{x_1} - \vec{y_1}) - (\vec{x_2} - \vec{y_2}) \rangle = 0 \tag{16}$$

The same equation appears in the derivation in (7). This holds iff $\vec{x_1} - \vec{y_1} = \vec{x_2} - \vec{y_2}$ (i.e., iff, by Definition 2, an analogy holds over $\{(x_1, y_1), (x_2, y_2)\}$)
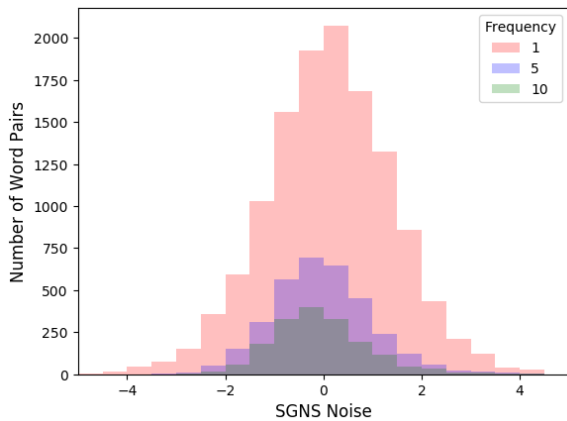
Figure 2: The noise distribution for an SGNS embedding model (i.e., $\langle \vec{x}, \vec{y_c} \rangle - [\mathrm{PMI}(x,y) - \log k]$) at various frequencies. The noise is normally distributed and the variance decreases as the frequency increases.
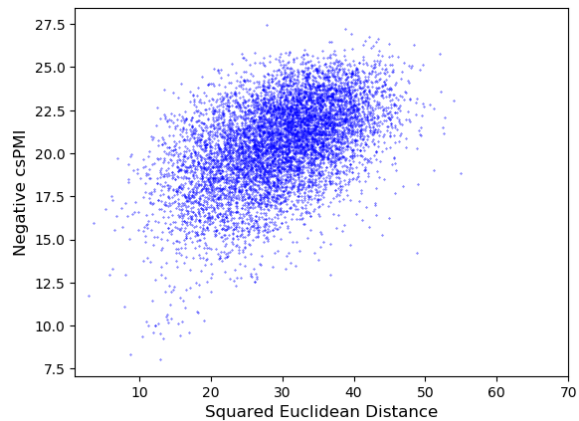

Figure 3: The negative csPMI for a word pair against the squared Euclidean distance between its SGNS word vectors. There is a positive correlation (Pearson's r = 0.502); the more similar two words are, the smaller the Euclidean distance between their vectors. In the normalized SGNS word space, the correlation is just as strong (Pearson's r = 0.514).

or if $\vec{w_c}$ is orthogonal to non-zero $(\vec{x_1} - \vec{y_1}) - (\vec{x_2} - \vec{y_2})$. Even if the context vector of some word is orthogonal to the difference between the relation vectors, not all are – as noted in section 3.4, not all word or context vectors lie in the same hyperplane in embedding space. Therefore, a linear word analogy holds over $\{(x_1, y_1), (x_2, y_2)\}$ iff for every word w, $p(w|x_1)/p(w|y_1) = p(w|x_2)/p(w|y_2)$. If this applies to every $(x_1, y_1), (x_2, y_2) \in S$, as stated in the conjecture, then the same analogy holds over S.

5 Experiments

Measuring Noise We uniformly sample word pairs in Wikipedia and estimate the noise (i.e., $\langle \vec{x}, \vec{y_c} \rangle - [\mathrm{PMI}(x,y) - \log k]$) using SGNS vectors trained on the same corpus. As seen in Figure 2, the noise has an approximately zero-centered Gaussian distribution and the variance of the noise is lower at higher frequencies, supporting our assumptions in section 3.5. As previously mentioned, this is partly why linear word analogies are robust to noise: in practice, they typically hold over very frequent word pairs, and at high frequencies, the amount of noise is often negligible.

Estimating csPMI The csPMI Theorem implies that if an analogy holds exactly over a set of word pairs when there is no reconstruction error, then each word pair has the same csPMI value. In Table 1, we provide the mean csPMI values for various analogies in Mikolov et al. (2013a) over the set of word pairs for which they should hold (e.g., $\{(Paris, France), (Berlin, Germany)\}$ for *capital-*

world). We also provide the accuracy of the vector arithmetic solutions for each analogy, found by minimizing cosine distance over the 100K most frequent words in the vocabulary.

As expected, when the variance in csPMI is lower, solutions to word analogies are more accurate: the Pearson correlation between accuracy and csPMI variance is -0.70 and statistically significant at the 1% level. This is because an analogy is more likely to hold over a set of word pairs when the displacement vectors are identical, and thus when the csPMI values are identical. Similar analogies, such as *capital-world* and *capital-common-countries*, also have similar mean csPMI values – our theory implies this, since similar analogies have similar displacement vectors. As the csPMI changes, the type of analogy gradually changes from geography (*capital-world*, *city-in-state*) to verb tense (*gram5-present-participle*, *gram7-past-tense*) to adjectives (*gram2-opposite*, *gram4-superlative*). We do not witness a similar gradation with the mean PMI, implying that analogies correspond uniquely to csPMI but not PMI.

Euclidean Distance Because the sum of two word vectors is not in the vocabulary, we cannot calculate co-occurrence statistics involving the sum, precluding us from testing Corollaries 1 and 2. We test Corollary 3 by uniformly sampling word pairs and plotting, in Figure 3, the negative csPMI against the squared Euclidean distance between the SGNS word vectors. As ex-

Analogy	Mean csPMI	Mean PMI	Median Word Pair Frequency	csPMI Variance	Accuracy
capital-world	−9.294	6.103	980.0	0.496	0.932
capital-common-countries	−9.818	4.339	3436.5	0.345	0.954
city-in-state	−10.127	4.003	4483.0	2.979	0.744
gram6-nationality-adjective	−10.691	3.733	3147.0	1.651	0.918
family	−11.163	4.111	1855.0	2.897	0.836
gram8-plural	−11.787	4.208	342.5	0.590	0.877
gram5-present-participle	−14.530	2.416	334.0	2.969	0.663
gram9-plural-verbs	−14.688	2.409	180.0	2.140	0.740
gram7-past-tense	−14.840	1.006	444.0	1.022	0.651
gram3-comparative	−15.111	1.894	194.5	1.160	0.872
gram2-opposite	−15.630	2.897	49.0	3.003	0.554
gram4-superlative	−15.632	2.015	100.5	2.693	0.757
currency	−15.900	3.025	19.0	4.008	0.092
gram1-adjective-to-adverb	−17.497	1.113	46.0	1.991	0.500

Table 1: The mean csPMI for analogies in Mikolov et al. (2013a) over the word pairs for which they should hold (e.g., *(Paris, France)* for *capital-world*). Similar analogies have a similar mean csPMI and arithmetic solutions are less accurate when the csPMI variance is higher (Pearson's $r = -0.70$). The type of analogy gradually changes with the csPMI, from geography (*capital-world*) to verb tense (*gram7-past-tense*) to adjectives (*gram2-opposite*).

pected, there is a moderately strong positive correlation (Pearson's $r = 0.502$): the more similar two words are (as measured by csPMI), the smaller the Euclidean distance between them in embedding space. The correlation is just as strong in the normalized SGNS word space, where Pearson's $r = 0.514$. As mentioned earlier, our assumption in section 3.4 that $p(w, w) \propto p(w)$ is justified because there is a strong positive correlation between the two (Pearson's $r = 0.825$).

Unsolvability   The csPMI Theorem reveals two reasons why an analogy may be unsolvable in a given embedding space: polysemy and corpus bias. Consider senses $\{x_1, ..., x_M\}$ of a polysemous word x. Assuming perfect reconstruction, a linear analogy f whose displacement has csPMI γ does not hold over (x, y) if $\gamma \neq \text{PMI}(x, y) + \log p(x, y) = \log[p(x_1|y) + ... + p(x_M|y)] p(y|x)$. The Theorem applies over all the senses of x, even if only a particular sense is relevant to the analogy. For example, while *(open,closed)::(high,low)* makes intuitive sense, it is unlikely to hold in practice, given that all four words are highly polysemous.

Even if *(a,b)::(x,y)* is intuitive, there is also no guarantee that $\text{csPMI}(a, b) \approx \text{csPMI}(x, y)$ and $\text{csPMI}(a, x) \approx \text{csPMI}(b, y)$ for a given training corpus. The less frequent a word pair, the more sensitive its csPMI to even small changes in frequency. Infrequent word pairs are also associated with more reconstruction error (see section 3.5), making it even more unlikely that the analogy will hold in practice. This is why the accuracy for the *currency* analogy is so low (see Table 1) – in

Wikipedia, currencies and their country co-occur with a median frequency of only 19.

6   Conclusion

In this paper, we explained why word analogies can be solved using vector arithmetic. We proved that an analogy holds in an SGNS or GloVe embedding space with no reconstruction error iff the co-occurrence shifted PMI is the same for every word pair and across any two word pairs, provided the row vectors of those words in the factorized word-context matrix are coplanar. This had three implications. First, we provided a formal proof of the Pennington et al. (2014) conjecture, the intuitive explanation of this phenomenon. Second, we provided novel justification for the addition of SGNS word vectors by showing that it automatically down-weights the more frequent word, as weighting schemes do *ad hoc*. Third, we provided the first rigorous explanation of why the Euclidean distance between word vectors is a good proxy for word dissimilarity. Most importantly, we provided empirical support of our theory and avoided making the strong assumptions in past work, making our theory a much more tenable explanation.

References

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. *arXiv preprint arXiv:1901.09813*.

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.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167. ACM.

Kawin Ethayarajh. 2018. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

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

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Tomas Mikolov, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878.

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.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520.

Douglas LT Rohde, Laura M Gonnerman, and David C Plaut. 2006. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8(627-633):116.

Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, pages 894–905.