What the Vec?
Towards Probabilistically Grounded Embeddings

Carl Allen **Ivana Balažević** **Timothy Hospedales**
School of Informatics
University of Edinburgh
{carl.allen, ivana.balazevic, t.hospedales}@ed.ac.uk

Abstract

Word2Vec (W2V) and Glove are popular word embedding algorithms that perform well on a variety of natural language processing tasks. The algorithms are fast, efficient and their embeddings widely used. Moreover, the W2V algorithm has recently been adopted in the field of graph embedding, where it underpins several leading algorithms. However, despite their ubiquity and the relative simplicity of their common architecture, *what* the embedding parameters of W2V and Glove learn and *why* that it useful in downstream tasks largely remains a mystery. We show that different interactions of *PMI vectors* encode semantic properties that can be captured in low dimensional word embeddings by suitable projection, theoretically explaining why the embeddings of W2V and Glove work, and, in turn, revealing an interesting mathematical interconnection between the semantic relationships of relatedness, similarity, paraphrase and analogy.

1 Introduction

Word2Vec[1] (W2V) [25] and Glove [29] are fast, straightforward algorithms for generating *word embeddings*, or vector representations of words, often considered points in a *semantic space*. Their embeddings perform well on downstream tasks, such as identifying word similarity by vector comparison (e.g. cosine similarity), and solving analogies, such as the well known "*man* is to *woman* as *king* is to *queen*", by the addition and subtraction of respective embeddings [26, 27, 18].

In addition, the W2V algorithm has recently been adopted within the growing field of *graph embedding*, where typically the aim is to represent each node in a common latent space such that their relative positioning can be used to predict edge relationships. Several state-of-the-art models for graph representation incorporate the W2V algorithm to learn node representations based on random walks over the graph [12, 30, 31]. Separately, word embeddings often underpin those of word sequences, such as sentence embeddings. Sentence embedding models can be complex [7, 16] but have recently been shown to sometimes learn little beyond the information available in word embeddings [38].

Despite their relative ubiquity, much remains unknown of the W2V and Glove algorithms, perhaps most fundamentally: (1) *what is learned* in the embedding parameters; and (2) *why that is useful* in downstream tasks. The latter is particularly interesting given the algorithms are unsupervised. To answer such core questions is of interest in itself, but may also lead to improved embedding algorithms, or enable better use to be made of the embeddings we have. For example, both algorithms generate two embedding matrices, but little is known of how they relate or should interact. Typically one matrix is simply discarded, while heuristics suggest that their mean can perform well [29] and elsewhere they are explicitly assumed identical [13, 3]. As for embedding interactions, a variety of heuristics have been proposed, e.g. cosine similarity [26] and *3CosMult* [18].

[1] We refer exclusively, throughout, to the more common implementation *Skipgram* with negative sampling.

Preprint. Under review.

Of works that seek to theoretically explain these embedding models [19, 13, 3, 8, 17], Levy and Goldberg [19] identify the loss function implicitly minimised by W2V and, thereby, the relationship between its word embeddings and *Pointwise Mutual Information* (PMI) of word co-occurrences. More recently, Allen and Hospedales [2] showed that this relationship underpins the linear interaction between embeddings of analogies. Building on these two results, our key contributions are:

- to show that semantic *similarity* is captured by high dimensional *PMI vectors* and, by considering geometric and probabilistic aspects of such vectors and their domain, to establish a hierarchical mathematical interrelationship between relatedness, similarity, paraphrases and analogies (Fig 2);
- to show that these semantic properties arise through *additive* interactions and so are best captured in low dimensional word embeddings by *linear* projection, thus explaining, by comparison of their loss functions, the presence of semantic properties in the embeddings of W2V and Glove;
- to derive a relationship between learned embedding matrices, proving that they necessarily differ (in the real domain), justifying the heuristic use of their mean, showing that different interactions are required to extract different semantic information, and enabling popular embedding comparisons, such as cosine similarity, to be semantically interpreted.

2 Background

The **Word2Vec** algorithm [25, 26] considers D word pairs $\{(w_{i_r}, c_{j_r})\}_{r=1}^{D}$ generated from a typically large text corpus. *Target* word w_i ranges over the corpus and *context* word c_j ranges over a window of size l, symmetric about w_i. For each observed word pair (*positive sample*), k random word pairs (*negative samples*) are drawn according to unigram distributions. For a chosen embedding dimensionality d and a dictionary $\mathcal{E}$ of n unique corpus words, the 2-layer neural network model applies the logistic sigmoid function to the product of two weight matrices $\mathbf{W}, \mathbf{C} \in \mathbb{R}^{d\times n}$ (by convention, $\mathbf{W}$ relates to target words). Columns of $\mathbf{W}$ and $\mathbf{C}$ are *word embeddings* for $\mathcal{E}$: the i^{th} column of $\mathbf{W}$, $\mathbf{w}_i \in \mathbb{R}^d$, represents the i^{th} word of $\mathcal{E}$ observed as the target word, denoted w_i; and $\mathbf{c}_j \in \mathbb{R}^d$, the j^{th} column of $\mathbf{C}$, represents the j^{th} word observed as a context word c_j.

Levy and Goldberg [19] show that the loss function minimised by the W2V algorithm is given by:

$$\ell_{W2V} = -\sum_{i=1}^{n}\sum_{j=1}^{n} \#(w_i, c_j)\log\sigma(\mathbf{w}_i^\top \mathbf{c}_j) + \tfrac{1}{D}\#(w_i)\#(c_j)\log(\sigma(-\mathbf{w}_i^\top \mathbf{c}_j)), \tag{1}$$

which is minimised if $\mathbf{w}_i^\top \mathbf{c}_j = \text{PMI}(w_i, c_j) - \log k$, where $\text{PMI}(w_i, c_j) = \log \frac{p(w_i, c_j)}{p(w_i)p(c_j)}$ is *Pointwise Mutual Information*. In matrix form, this equates to factorisation of a *shifted* PMI matrix (**SPMI**):

$$\mathbf{W}^\top \mathbf{C} \;=\; \mathbf{SPMI} \;\in \mathbb{R}^{n\times n}\,. \tag{2}$$

Glove [29] has the same architecture as W2V, but a different loss function, minimised when:

$$\mathbf{w}_i^\top \mathbf{c}_j = \log p(w_i, c_j) - b_i - b_j + \log Z, \tag{3}$$

for biases b_i, b_j and normalising constant Z. In principle, the biases provide flexibility, expanding the family of statistical relationships that Glove embeddings can learn.

Analogies are word relationships, such as the canonical "*man* is to *woman* as *king* is to *queen*", that pique interest because their respective word embeddings are observed to form linear relationships [27, 18]. It was recently shown [2] that this phenomenon follows from underlying properties of *PMI vectors*, or rows of the *unshifted* PMI matrix (**PMI** $\in \mathbb{R}^{n\times n}$) for each word. In doing so, the authors motivate a probabilistic definition of the *paraphrasing* of a set of words $\mathcal{W}$ by a word w_* as being when the *induced distributions* of w_* and $\mathcal{W}$ are similar, where an induced distribution is a conditional probability distribution $p(\mathcal{E}|\circ)$ over all words in $\mathcal{E}$ given a particular observation $\circ$, and comparison is by the *Kullback–Leibler (KL) divergence* (recall $\text{KL}(p \,\|\, q) = \sum_i p_i \log p_i/q_i$).

3 Related Work

Many works explore empirical properties of word embeddings (e.g. [18, 22, 4]). We focus here on those that set out to theoretically explain why the *W2V* and *Glove* embedding algorithms succeed in producing word embeddings that capture semantic properties useful in downstream tasks. The first of these is the mentioned derivation by Levy and Goldberg [19] of the loss function (1) and the relationship to PMI that minimises it (2). Generative language models are proposed by Hashimoto et al. [13] and Arora et al. [3] to explain the structure found in word embeddings. However, each contains strong *a priori* assumptions of an underlying geometry that we do not require (in fact, it can be seen that several assumptions of [3] fail in practice (Appendix D)). Cotterell et al. [8] and Landgraf and Bellay [17] show that W2V's loss function performs *exponential (binomial) PCA* [6], however that follows from the *binomial* negative sampling and thus describes the algorithm's mechanics, not *why* it works. Several works focus on the observed linearity of analogical embeddings [3, 11, 2, 9], but, other than [2] as described, fail to rigorously link semantics to embedding geometry.

To our knowledge, no previous work explains how all of the semantic properties of relatedness, similarity, paraphrasing and analogies are encoded in PMI vectors and how they then translate into the semantic properties observed in low dimensional word embeddings of W2V and Glove.

4 PMI: linking geometry to semantics

We first briefly consider the derivative of W2V's loss function (1) with respect to embedding $\mathbf{w}_i$:

$$\tfrac{1}{D}\nabla_{\mathbf{w}_i}\ell_{W2V} = \sum_{j=1}^{n}\Big(\underbrace{p(w_i,c_j)+kp(w_i)p(c_j)}_{\mathbf{p}_j^{(i)}}\Big)\Big(\underbrace{\sigma(\text{SPMI}(w_i,c_j)) - \sigma(\mathbf{w}_i^\top\mathbf{c}_j)}_{\mathbf{e}_j^{(i)}}\Big)\mathbf{c}_j = \mathbf{C}\,\mathbf{P}^{(i)}\mathbf{e}^{(i)} \tag{4}$$

for diagonal matrix $\mathbf{P}^{(i)} = diag(\mathbf{p}^{(i)}) \in \mathbb{R}^{n\times n}$; vectors $\mathbf{p}^{(i)}, \mathbf{e}^{(i)} \in \mathbb{R}^n$ of probability and error terms, respectively; and count-based empirical probability estimates $p(\cdot)$, e.g. $p(w_i) = \#(w_i)/D$. This shows that the loss function (1) is indeed minimised if (2) holds. However, that requires **W**, **C** to have rank at least that of **SPMI**. Otherwise, as for typical word embeddings with $d \ll n$, the loss function is minimised if probabilistically weighted error terms, diag$(\mathbf{p}^{(i)})\,\mathbf{e}^{(i)}$, are orthogonal to the rows of **C**. That is, the loss function serves as a *non-linear projection* (due to $\sigma(\cdot)$ in $\mathbf{e}^{(i)}$) of rows of the **SPMI** matrix onto rows of **C**. (The distinction between the roles of **W** and **C** is maintained throughout but arbitrary, projection onto rows of **W** could be considered equally.)

Viewing word embeddings as low-dimensional projections of rows of **PMI**, or *PMI vectors*, we consider their properties, their domain and interactions to understand why their projections are useful (denoting matrix rows by superscript, $\mathbf{PMI}^i$ is the PMI vector of w_i). Our aim is to then choose a loss function that preserves desirable properties while approximating the low-rank factorisation:

$$\mathbf{w}_i^\top\mathbf{c}_j \approx \text{PMI}(w_i,c_j) \qquad \text{or in matrix form} \qquad \mathbf{W}^\top\mathbf{C} \approx \mathbf{PMI}. \tag{5}$$

Note there is no *shift* of (2), which can be seen as an artefact of the W2V algorithm (see Appendix B).

4.1 The domain of PMI vectors

$\mathbf{PMI}^i$ has components $\text{PMI}(w_i,c_j) \in \mathbb{R}$, for all context words $c_j \in \mathcal{E}$, which can be seen as log ratios of conditional to marginal probabilities:

$$\text{PMI}(w_i,c_j) = \log\tfrac{p(c_j,w_i)}{p(w_i)p(c_j)} = \log\tfrac{p(c_j|w_i)}{p(c_j)}. \tag{6}$$

Any (log) change in probability of observing c_j, having observed w_i relative to if no such observation occurred, can be thought of as *due* to w_i. Thus $\text{PMI}(w_i,c_j)$ captures dependence, or influence, of one word on another. Specifically, by reference to marginal probability $p(c_j)$, $\text{PMI}(w_i,c_j) > 0$ implies c_j is more likely to occur in the presence of w_i; $\text{PMI}(w_i,c_j) < 0$ implies c_j is less likely to occur given w_i; and $\text{PMI}(w_i,c_j) = 0$ indicates that w_i and c_j occur independently, with no impact on one another, i.e. they are unrelated. A single PMI value thus reflects the semantic property of word *relatedness*, as previously noted [36, 5, 14]. A PMI vector therefore reflects change in $p(\mathcal{E})$, the probability distribution over all words, given (or due to) w_i:

$$\mathbf{PMI}^i = \big\{\log\tfrac{p(c_j|w_i)}{p(c_j)}\big\}_{c_j\in\mathcal{E}} = \log\tfrac{p(\mathcal{E}|w_i)}{p(\mathcal{E})}. \tag{7}$$

While PMI values are unconstrained in $\mathbb{R}$, PMI vectors are constrained to an n−1 dimensional surface $\mathcal{S} \subset \mathbb{R}^n$, where each dimension corresponds to a word (Fig 1)

- the vector of numerator terms $\mathbf{q}\!=\!p(\mathcal{E}|w_i)$ lies on the simplex $\mathcal{Q}\!\subset\!\mathbb{R}^n$;
- dividing all $\mathbf{q}\!\in\!\mathcal{Q}$ by $p(\mathcal{E})\in\mathcal{Q}$, gives probability ratio vectors $\frac{p(\mathcal{E}|w_i)}{p(\mathcal{E})}$ that define a "stretched simplex" $\mathcal{R}\!\subset\!\mathbb{R}^n$ passing through $\mathbf{1}\!\in\!\mathbb{R}^n$ that, for all $c_j\!\in\!\mathcal{E}$, has a vertex on axis j at $\frac{1}{p(c_j)}$; and
- the natural logarithm transforms $\mathcal{R}$ to a curved surface $\mathcal{S}$, such that $\mathbf{PMI}^i\!\in\!\mathcal{S},\ \forall w_i\!\in\!\mathcal{E}$.

Note $p(\mathcal{E})$ uniquely determines $\mathcal{S}$. Considering point $\mathbf{s}\!=\!\log\frac{\mathbf{q}}{\mathbf{p}}\!\in\!\mathcal{S}$ as an element-wise log ratio of probability vectors $\mathbf{p}\!=\!p(\mathcal{E}), \mathbf{q}\!\in\!\mathcal{Q}$, shows $\mathcal{S}$ has the following properties (proofs in Appendix A):

1. **$\mathcal{S}$, and any subsurface of $\mathcal{S}$, is non-linear.** PMI vectors are therefore not constrained to a linear subspace, identifiable by low-rank factorisation of the PMI matrix, as may seem suggested by the association of (5) with *Principal Component Analysis* (PCA) [15, 35].
2. **$\mathcal{S}$ contains the origin $\mathbf{0}\in\mathbb{R}^n$,** which can be considered a PMI vector of the *null word* $\emptyset$, i.e. $\mathbf{PMI}^{\emptyset} = \log\frac{p(\mathcal{E}|\emptyset)}{p(\mathcal{E})} = \log\frac{p(\mathcal{E})}{p(\mathcal{E})} = \mathbf{0}$, as contemplated by [9].
3. **Probability vector $\mathbf{q}$ is normal to the tangent plane of $\mathcal{S}$** at $\mathbf{s} = \log\mathbf{q}/p(\mathcal{E})\!\in\!\mathcal{S}$.
4. **$\mathcal{S}$ does not intersect with the fully positive or fully negative orthants** (excluding $\mathbf{0}$). Thus PMI vectors are not *isotropically* (i.e. uniformly) distributed in space (as assumed in [3]).
5. **The sum of 2 points $\mathbf{s}+\mathbf{s}'$ lies in $\mathcal{S}$ only for certain $\mathbf{s},\ \mathbf{s}'\!\in\!\mathcal{S}$.** That is, for any $\mathbf{s}\!\in\!\mathcal{S}$ ($\mathbf{s}\!\neq\!\mathbf{0}$), there exists a (strict) subset $\mathcal{S}_s\!\subset\!\mathcal{S}$, such that $\mathbf{s}+\mathbf{s}'\!\in\!\mathcal{S}$ *iff* $\mathbf{s}'\!\in\!\mathcal{S}_s$. Trivially $\mathbf{0}\!\in\!\mathcal{S}_s,\ \forall\mathbf{s}\!\in\!\mathcal{S}$.

Note that while all PMI vectors lie in $\mathcal{S}$, certainly not all (infinite) points in $\mathcal{S}$ correspond to the (finite) PMI vectors of words. Interestingly, points 2 and 5 allude to properties of a *vector space*, often the desired structure for a *semantic space* [13]. Whilst the domain of PMI vectors is clearly not a vector space, addition and subtraction of PMI vectors do have *semantic meaning*, as we now show.

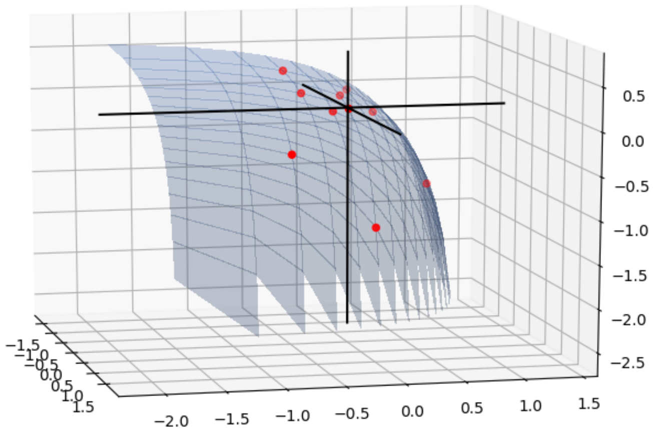


Figure 1: The PMI surface $\mathcal{S}$, showing sample PMI vectors of words

4.2 Subtraction of PMI vectors finds similarity

Taking the definition of paraphrasing (S.2), we consider a word w_i that paraphrases a word set $\mathcal{W}$ containing only a single word, e.g. $\mathcal{W}\!=\!\{w_j\}$. Since paraphrasing requires induced distributions to be similar, this intuitively finds words that are *interchangeable* – in the limit, w_j itself or, less trivially, a synonym, i.e. w_i and w_j must be *similar*. Thus, similarity translates to a low KL divergence between $p(\mathcal{E}|w_i)$ and $p(\mathcal{E}|w_j)$. Interestingly, the difference between two PMI vectors is given by:

$$\boldsymbol{\rho}^{i,j} = \mathbf{PMI}^i - \mathbf{PMI}^j = \log\frac{p(\mathcal{E}|w_i)}{p(\mathcal{E}|w_j)}, \tag{8}$$

a vector of un-weighted KL divergence components. Thus, if dimensions were suitably weighted, the sum of difference components (similar to Manhattan distance but *directed*) would equate to a KL divergence between induced distributions. Equivalently, the vector difference $\boldsymbol{\rho}^{i,j}$ gives a KL divergence as the dot product $\mathbf{q}_i^\top\boldsymbol{\rho}^{i,j}$ with probability vector $\mathbf{q}_i = p(\mathcal{E}|w_i)$. From point 3, $\mathbf{q}_i$ is in fact the normal to the surface $\mathcal{S}$ at $\mathbf{PMI}^i$ (with unit l_1 norm). Taking -$\mathbf{q}_j^\top\boldsymbol{\rho}^{i,j}$, i.e. projecting onto the normal to $\mathcal{S}$ at $\mathbf{PMI}^j$, gives the alternate KL divergence. (Intuition for the semantic meaning captured by each KL divergence is discussed in [2].).

4.3 Addition of PMI vectors finds paraphrases

From geometric arguments (point 5), we know that only certain pairs of points in $\mathcal{S}$ sum to another point in the surface. We can also consider *probabilistic* conditions for PMI vectors to sum to another:

$$\begin{aligned}\mathbf{PMI}^i + \mathbf{PMI}^j &= \log \tfrac{p(\mathcal{E}|w_i)}{p(\mathcal{E})} + \log \tfrac{p(\mathcal{E}|w_j)}{p(\mathcal{E})} = \log \tfrac{p(w_i|\mathcal{E})\,p(w_j|\mathcal{E})}{p(w_i)\,p(w_j)} \\ &= \underbrace{\log \tfrac{p(w_i,w_j|\mathcal{E})}{p(w_i,w_j)}}_{\mathbf{PMI}^{i,j}} - \underbrace{\log \tfrac{p(w_i,w_j|\mathcal{E})}{p(w_i|\mathcal{E})p(w_j|\mathcal{E})}}_{\boldsymbol{\sigma}^{ij}} + \underbrace{\log \tfrac{p(w_i,w_j)}{p(w_i)p(w_j)}}_{\tau^{ij}} = \mathbf{PMI}^{i,j} - \boldsymbol{\sigma}^{ij} + \tau^{ij}\mathbf{1}, \quad (9)\end{aligned}$$

where, overloading notation, $\mathbf{PMI}^{i,j}$ is a vector of PMI terms involving $p(\mathcal{E}|w_i, w_j)$, the distribution induced by w_i *and* w_j observed together;[2] and $\boldsymbol{\sigma}^{ij} \in \mathbb{R}^n$, $\tau^{ij} \in \mathbb{R}$ (borrowed from [2]) are conditional and marginal dependence terms, as indicated. This shows that $\mathbf{s} = \mathbf{PMI}^i + \mathbf{PMI}^j$ lies in $\mathcal{S}$ *if* words w_i, $w_j \in \mathcal{E}$ occur both *independently and conditionally independently* given each and every word in $\mathcal{E}$. If so, $\mathbf{s} = \mathbf{PMI}^{i,j}$ corresponding to the joint occurrence of w_i and w_j, and (from point 5) $\mathbf{PMI}^j \in \mathcal{S}_{\mathbf{PMI}^i}$ and $\mathbf{PMI}^i \in \mathcal{S}_{\mathbf{PMI}^j}$. If not, error vector $\boldsymbol{\varepsilon}^{ij} = \boldsymbol{\sigma}^{ij} - \tau^{ij}\mathbf{1}$ separates $\mathbf{s}$ from $\mathbf{PMI}^{i,j}$ (and $\mathbf{s} \in \mathcal{S}$ only if by meaningless coincidence). Note that probabilistic aspects here mirror [2], but we look to combine this with our geometric understanding. Considering $\mathbf{PMI}^{i,j}$, whilst certainly in $\mathcal{S}$, the extent to which it relates to $\mathbf{PMI}^k$ of some (single) word w_k depends on paraphrasing, i.e. the similarity between distributions induced by w_k and $\{w_i, w_j\}$, as captured by $\boldsymbol{\rho}^{ij} = \mathbf{PMI}^k - \mathbf{PMI}^{i,j}$. Thus the "gap" between $\mathbf{PMI}^i + \mathbf{PMI}^j$ and $\mathbf{PMI}^k$ can be geometrically interpreted in terms of components $\boldsymbol{\varepsilon}^{ij}$ (dependence) that moves *to* the surface $\mathcal{S}$, and $\boldsymbol{\rho}^{ij}$ (paraphrase) that moves *along* it. In a sense, the latter is the interesting relationship that the former (independent of w_k) potentially obscures. (For further joint implication of geometric and probabilistic considerations see Appendix C.)

4.4 Linear combinations of PMI vectors find analogies

PMI vectors of analogical relationships "w_a is to w_{a^*} as w_b is to w_{b^*}" have been shown [2] to satisfy:

$$\mathbf{PMI}^{b^*} \;\approx\; \mathbf{PMI}^{a^*} - \mathbf{PMI}^{a} + \mathbf{PMI}^{b}. \quad (10)$$

A more extensive argument based on paraphrasing, with error terms similar to those in Section 4.3, compares PMI vectors of pairs of words, giving $\mathbf{PMI}^a + \mathbf{PMI}^{b^*} \approx \mathbf{PMI}^{a^*} + \mathbf{PMI}^b$ and thus (10).

5 Encoding PMI: from PMI vectors to word embeddings

Having established how semantic properties, desirable in word embeddings, arise in high dimensional PMI vectors, we consider how they can be transferred to low dimensional representations. A key observation is that all PMI vector interactions, for similarity (8), paraphrases (9) and analogies (10), are *additive*, and therefore preserved under *linear* projection. By comparison, the loss function of W2V (1) projects PMI vectors non-linearly, and that of Glove (3) does project linearly, but not necessarily PMI vectors. Linear projection can be achieved by the least squares loss function:

$$\ell_{LSQ} \;=\; \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big(\mathbf{w}_i{}^\top\mathbf{c}_j - \mathrm{PMI}(w_i, c_j)\big)^2. \quad (11)$$

ℓ_{LSQ} is minimised when $\nabla_{\mathbf{W}^\top}\ell_{LSQ} = (\mathbf{W}^\top\mathbf{C} - \mathbf{PMI})\mathbf{C}^\top = 0$, or $\mathbf{W}^\top = \mathbf{PMI}\,\mathbf{C}^\dagger$, for $\mathbf{C}^\dagger = \mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1}$ the *Moore–Penrose pseudoinverse* of $\mathbf{C}$. This explicit linear projection allows interactions performed between word embeddings, e.g. dot product, to be mapped to interactions between PMI vectors, and thereby semantically interpreted. However, we do better still by understanding how $\mathbf{W}$ and $\mathbf{C}$ relate.

5.1 The relationship between W and C

Whilst W2V and Glove train two embedding matrices, typically only $\mathbf{W}$ is used and $\mathbf{C}$ discarded. Thus, although relationships are learned between $\mathbf{W}$ and $\mathbf{C}$, they are tested between $\mathbf{W}$ and $\mathbf{W}$. If

[2] Note that w_i, w_j are both target words but, by symmetry, we can interchange roles of context and target words to compute $p(\mathcal{E}|w, w')$ based on the distribution of target words when w_i and w_j are both context words.

$\mathbf{C}=\mathbf{W}$ the distinction falls away, but that is not found to be the case in practice. Here we consider why it might be that $\mathbf{W}\neq\mathbf{C}$ and, if so, what relationship between $\mathbf{W}$ and $\mathbf{C}$ does exist.

If the symmetric matrix **PMI** were *positive semi-definite* (PSD), its closest low-rank approximation (in an l_2 sense, minimising ℓ_{LSQ}) would be given by its real eigendecomposition $\mathbf{PMI}=\boldsymbol{\Pi}\boldsymbol{\Lambda}\boldsymbol{\Pi}^\top$, $\boldsymbol{\Pi},\boldsymbol{\Lambda}\in\mathbb{R}^{n\times n}$, $\boldsymbol{\Pi}^\top\boldsymbol{\Pi}=\mathbf{I}$. Specifically, ℓ_{LSQ} is minimised by $\mathbf{W}=\mathbf{C}=\mathbf{S}^{1/2}\mathbf{U}^\top$, where $\mathbf{S}\in\mathbb{R}^{d\times d}$, $\mathbf{U}\in\mathbb{R}^{d\times n}$ are $\boldsymbol{\Lambda}$, $\boldsymbol{\Pi}$, respectively, truncated to their d largest eigenvalue components, giving $\mathbf{W}^\top\mathbf{W}\approx\mathbf{PMI}$. Indeed, many matrix pairs minimise ℓ_{LSQ}, as constructed by multiplying $\mathbf{W}$, $\mathbf{C}$, respectively, by any invertible $\mathbf{M}\in\mathbb{R}^{d\times d}$ and its inverse $\mathbf{M}^{-1}$; but, of these, $\mathbf{W}$, $\mathbf{C}$ are unique, up to rotation and permutation, in satisfying $\mathbf{W}=\mathbf{C}$. This is ideal for word embeddings as the number of learned parameters is halved and consideration of whether to use $\mathbf{W}$, $\mathbf{C}$ or both falls away.

PMI is not typically PSD in practice, however, and no such ideal *real* factorisation exists, i.e. since **PMI** has negative eigenvalues, $\mathbf{S}^{1/2}$ is complex and any embeddings minimising ℓ_{LSQ} with $\mathbf{W}=\mathbf{C}$ must also be complex, i.e. $\mathbf{W}\in\mathbb{C}^{d\times n}$. Complex word embeddings have been considered, e.g. [23, 21], and may be of interest to explore, however the word embeddings we aim to explain are real so we keep to the real domain. We have thus established that $\mathbf{W},\mathbf{C}\in\mathbb{R}^{d\times n}$ *cannot be equal*, which we note contradicts the assumption $\mathbf{W}=\mathbf{C}$ sometimes made [13, 3]. Returning to the eigendecomposition, where $\mathbf{S}$ contains the d largest *absolute* eigenvalues (which may be negative) and $\mathbf{U}$ the corresponding eigenvectors, we factor $\mathbf{S}$ as $\mathbf{S}=\mathbf{S}'\mathbf{I}'$ where $\mathbf{S}'=|\mathbf{S}|$ and $\mathbf{I}'=$ *sign*$(\mathbf{S})$, $\mathbf{I}'_{ii}\in\{1,\text{–}1\}$. Choosing $\mathbf{W}=\mathbf{S}'^{1/2}\mathbf{U}^\top$gives $\mathbf{W}^\top\mathbf{I}'\mathbf{W}\approx\mathbf{PMI}$, and $\mathbf{C}=\mathbf{I}'\mathbf{W}$ recovers $\mathbf{W}^\top\mathbf{C}\approx\mathbf{PMI}$, with $\mathbf{W}\neq\mathbf{C}$ but $\mathbf{W}^i=\pm\mathbf{C}^i$ for rows of $\mathbf{W}$, $\mathbf{C}$ (recall word embeddings are columns). $\mathbf{W},\mathbf{C}$ can be seen as *quasi*-complex conjugate. This mirrors PCA of the PMI matrix, but makes the relationship between $\mathbf{W}$ and $\mathbf{C}$ explicit. This choice is most parameter efficient in the non-PSD case of typical **PMI** ($(n+1)d$ as opposed to $2nd$). As above, such $\mathbf{W},\mathbf{C}$ are among a family of matrix pairs ($\mathbf{M}^\top\mathbf{W},\mathbf{M}^{-1}\mathbf{C}$) that minimise ℓ_{LSQ}, thus the underlying relationship exists but may be obscured.

5.2 Interpreting embedding interactions

Diverse interactions between word embeddings have been used in the literature, e.g. cosine similarity [26] and *3CosMult* [18], with little theoretical analysis. Based on our theoretical findings, in particular $\mathbf{C}=\mathbf{I}'\mathbf{W}$, we semantically interpret commonly used interactions and assess implications of interacting embeddings of $\mathbf{W}$ with one another, not those of $\mathbf{C}$. Noting that $\mathbf{C}^\dagger=\mathbf{U}\mathbf{S}'^{-1/2}\mathbf{I}'$ and $\mathbf{W}^\top\mathbf{C}=\mathbf{U}\mathbf{S}\mathbf{U}^\top$, we define reconstruction error matrix $\mathbf{E}=\mathbf{PMI}\text{ – }\mathbf{W}^\top\mathbf{C}=\mathbf{V}\mathbf{T}\mathbf{V}^\top$, where $\mathbf{V}$, $\mathbf{T}$ contain components of $\boldsymbol{\Pi}$, $\boldsymbol{\Sigma}$ corresponding to the $n\text{–}d$ *smallest* absolute eigenvalues (as omitted from $\mathbf{U}$, $\mathbf{S}$); and $\mathbf{F}=\mathbf{U}(\frac{\mathbf{S}-\mathbf{S}'}{2})\mathbf{U}^\top$, correspondon to eigencomponents of **PMI** with *negative* eigenvalues. We define *mean embeddings* $\mathbf{a}_i$, as columns of $\mathbf{A}=\frac{\mathbf{W}+\mathbf{C}}{2}$, whereby $\mathbf{A}=\mathbf{U}\mathbf{S}'^{1/2}\mathbf{I}''$, with $\mathbf{I}''=\frac{\mathbf{I}+\mathbf{I}'}{2}$, $\mathbf{I}''_{ii}\in\{0{,}1\}$.

Dot Product: We compare the following interactions that pertain to predicting relatedness:

$$\begin{array}{rrcllcl}
\mathbf{W},\ \mathbf{C}: & \mathbf{w}_i^\top\mathbf{c}_j & = & \mathbf{U}^i\,\mathbf{S}\,\mathbf{U}^{j\top} & & = & \mathbf{PMI}_{i,j}-\mathbf{E}_{i,j}\\
\mathbf{W},\mathbf{W}: & \mathbf{w}_i^\top\mathbf{w}_j & = & \mathbf{U}^i\,\mathbf{S}'\,\mathbf{U}^{j\top} & =\ \mathbf{U}^i(\mathbf{S}-(\mathbf{S}-\mathbf{S}'))\mathbf{U}^{j\top} & = & \mathbf{PMI}_{i,j}-\mathbf{E}_{i,j}-2\,\mathbf{F}_{i,j}\\
\mathbf{A},\ \mathbf{A}: & \mathbf{a}_i^\top\mathbf{a}_j & = & \mathbf{U}^i\,\mathbf{S}'\,\mathbf{I}''\,\mathbf{U}^{j\top} & =\ \mathbf{U}^i(\mathbf{S}-(\frac{\mathbf{S}-\mathbf{S}'}{2}))\mathbf{U}^{j\top} & = & \mathbf{PMI}_{i,j}-\mathbf{E}_{i,j}-\ \ \mathbf{F}_{i,j}
\end{array}$$

Thus the dot product $\mathbf{w}_i^\top\mathbf{w}_j$ *overstates* the estimate of $\text{PMI}(w_i,w_j)$, and thus relatedness, given by $\mathbf{w}_i^\top\mathbf{c}_j=\mathbf{w}_i^\top\mathbf{I}'\mathbf{w}_j$, by twice any negative eigenvalue related component. Interestingly, that error is halved by mean embeddings.

Difference sum: $(\mathbf{w}_i-\mathbf{w}_j)^\top\mathbf{1}=(\mathbf{PMI}^i-\mathbf{PMI}^j)\mathbf{C}^\dagger\mathbf{1}=\log\frac{p(\mathcal{E}|w_i)}{p(\mathcal{E}|w_j)}(\mathbf{U}\mathbf{S}'^{-1/2}\mathbf{I}')\mathbf{1}$

Summing embedding difference components appears to correspond closely to KL divergence, and so *similarity*, more so than for PMI vectors (S.4.2) as dimensions are weighted, by $\mathbf{v}=\mathbf{U}\mathbf{S}'^{-1/2}\mathbf{I}'\mathbf{1}$. Unlike for true KL divergence, $\mathbf{v}$ is not normalised and is fixed irrespective of $\mathbf{w}_i,\mathbf{w}_j$. We speculate that $\mathbf{v}$ serves to down-weight low probability dimensions (encountered least in the dimensionality reduction process) and thereby“outliers” to which PMI is known to be sensitive [37], perhaps loosely reflecting KL divergences.

Euclidean distance: $\|\mathbf{w}_i-\mathbf{w}_j\|_2=\|(\log\frac{p(\mathcal{E}|w_i)}{p(\mathcal{E}|w_j)})\mathbf{C}^\dagger\|_2$ shows no obvious meaning.

Table 1: Accuracy in semantic tasks using different loss functions on the text8 corpus [24].

Model	Loss	Relationship	Relatedness[1]	Similarity[1]	Analogy[25]
$W2V$	W2V	$\mathbf{W}^\top\mathbf{C}\approx\mathbf{PMI}$	.628	.703	.283
$W{=}C$	LSQ	$\mathbf{W}^\top\mathbf{W}\approx\mathbf{PMI}$	.721	.786	.411
LSQ	LSQ	$\mathbf{W}^\top\mathbf{C}\approx\mathbf{PMI}$	**.727**	**.791**	**.425**

Cosine similarity: Surprisingly, the frequently used and effective $cosine(w_i, w_j) = \frac{\mathbf{w}_i^\top\mathbf{w}_j}{\|\mathbf{w}_i\|\|\mathbf{w}_j\|}$ shows no immediate meaning. However, we note that this interaction is often applied across multiple tasks, particularly relatedness and similarity [33, 4], which we have seen correspond to different interactions: dot product of $\mathbf{w}_i$ and $\mathbf{c}_j$, and sum of difference between $\mathbf{w}_i$ and $\mathbf{w}_j$. Thus, by combining a dot product with interactions of $\mathbf{w}_i, \mathbf{w}_j$, we conjecture that the popular, but previously unjustified, cosine similarity may serve as a blended measure of these the two semantic relationships.

Other statistical word embedding relationships assumed in [3] are considered in Appendix D.

6 Empirical evidence

Word embeddings (especially those of W2V) have been widely studied empirically, with many experimental findings. Here we draw on previous results and run test experiments to show empirical support for our main theoretical results:

1. Analogies form as linear relationships between linear projections of PMI vectors (S.4.4)

   Whilst previously explained [2], we emphasise that their rationale for this well known phenomenon fits precisely within our broader explanation of W2V and Glove embeddings. Further, re-ordering paraphrase questions is observed to materially affect prediction accuracy [22], which can be justified in terms of the explanation provided in [2] (see Appendix E).

2. Linear projection of *additive* PMI vectors captures semantic properties more accurately than the non-linear projection of W2V (S.5).

   Several works consider alternatives to the W2V loss function [19, 20]. However, not finding this direct comparison, we implement it ourselves (detail below). Isolating the impact of the loss function, comparison of models $W2V$ vs LSQ (Table 1) shows a material improvement from linear projection across all semantic tasks.

3. Word embedding matrices **W** and **C** are dissimilar (S.5.1).

   W, **C** are typically found to differ, e.g. [26, 29, 28]. To demonstrate the difference, we include a comparison tying $\mathbf{W}=\mathbf{C}$ in the previous experiment. Comparing models $W{=}C$ and LSQ (Table 1) shows a small but consistent improvement in the former despite a lower data to parameter ratio.

4. Dot products recover PMI (relatedness) with different accuracy: $\mathbf{w}_i^\top\mathbf{c}_j \geq \mathbf{a}_i^\top\mathbf{a}_j \geq \mathbf{w}_i^\top\mathbf{w}_j$ (S.5.2).

   The use of average embeddings $\mathbf{a}_i^\top\mathbf{a}_j$ over $\mathbf{w}_i^\top\mathbf{w}_j$ is a well-known heuristic [29, 20]. More recently, [4] show that relatedness correlates noticeably better to $\mathbf{w}_i^\top\mathbf{c}_j$ than either of those "symmetric" choices. Note that although we show the PMI approximation error is halved using $\mathbf{a}_i^\top\mathbf{a}_j$ (*cf* $\geq \mathbf{w}_i^\top\mathbf{w}_j$), prediction accuracy is not expected to change linearly.

5. Relatedness is reflected by interactions between **W** and **C** embeddings, similarity by interactions between **W** and **W**. (S.5.2)

   Asr et al. [4] compare human judgements of similarity and relatedness to cosine similarity between combinations of **W**, **C** and **A**. The authors find a "very consistent" support for their conclusion that "WC ... best measures ... relatedness" and "similarity [is] best predicted by ... WW". An example is given for *house*: $\mathbf{w}_i^\top\mathbf{w}_j$ gives *mansion*, *farmhouse* and *cottage*, i.e. similar or synonymous words; $\mathbf{w}_i^\top\mathbf{c}_j$ gives *barn*, *residence*, *estate*, *kitchen*, i.e. related words.

Models: As we perform a standard comparison of loss functions, similar to [19, 20], we leave experimental details to Appendix F. In summary, we learn 500 dimensional embeddings from word co-occurrences extracted from a standard corpus ("text8" [24]). We implement loss function (1) explicitly as model $W2V$. Models $W{=}C$ and LSQ use least squares loss (11), with constraint $\mathbf{W}=\mathbf{C}$ in the latter (see point 3 above). Evaluation on popular data sets [1, 25] uses the Gensim toolkit [32].

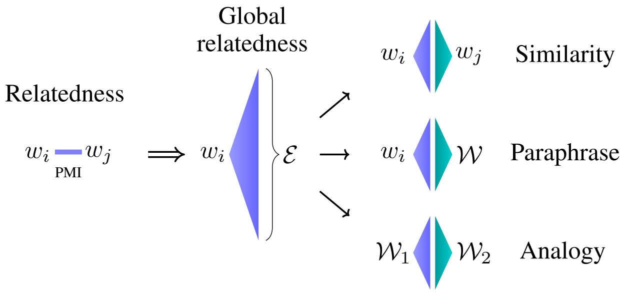


Figure 2: Interconnection between semantic relationships: relatedness is a base pairwise comparison (measured by PMI); *global relatedness* considers relatedness to all words (PMI vector); similarity, paraphrases and analogies depend on global relatedness between words ($w \in \mathcal{E}$) and word sets ($\mathcal{W} \subseteq \mathcal{E}$).

7 Discussion

Having considered how semantic structure is captured in word embeddings projected from PMI vectors, we can invert our perspective to see that this implies an interesting, mathematically defined hierarchical interplay between the semantic relationships themselves, that is, between relatedness, similarity, paraphrase and analogy (Fig 2). At the core is *relatedness*, which correlates with PMI, both empirically [36, 5, 14] and intuitively (S.4.2). As a pairwise comparison of words, relatedness acts somewhat akin to a *kernel* (note, an actual kernel requires **PMI** to be PSD), whereby words can be considered numerically according to their *relatedness to all words* (captured by a PMI vector). Instead of comparing words pair-wise, this enables words to be compared according to how they each relate to all other words, or how they both *globally relate*. Given this meta-comparison, we see that one word is *similar* to another if they are globally related (1-1); a *paraphrase* requires one word to globally relate to the joint occurrence of a set of words (1-n); and analogies arise when joint occurrences of word pairs are globally related (n-n). Taking the "kernel" theme further, the PMI matrix corresponds to a kernel matrix, and word embeddings to representations derived from *kernelised PCA* [34].

8 Conclusion

In this work, we take two main results, the well known relationship to PMI learned by W2V embeddings [19] and a recent connection drawn between PMI and analogies [2], to show that word embedding models that *linearly* project PMI vectors capture the semantic properties of relatedness, similarity, paraphrase and analogy. The loss functions of W2V (2) and Glove (3) can be seen to approximate such a projection, *non-linearly* in the case of W2V and not fully capturing PMI in Glove, thus explaining why their embeddings exhibit semantic properties useful in downstream tasks.

From linear projection, we derive a relationship between embedding matrices **W** and **C** that enables common word embedding interactions to be semantically interpreted. This suggests that the familiar cosine similarity may serve as a blend of more semantically meaningful interactions. Our theoretical results explain several empirical observations, e.g. why **W** and **C** are not found to be equal, despite representing the same words, their symmetric treatment in the loss function and a symmetric PMI matrix; why mean embeddings (**A**) are often found to outperform those from **W**; and why relatedness corresponds to interactions between **W** and **C**, and similarity to interactions between **W** and **W**.

We reveal an interesting hierarchical structure between these semantic properties in which *relatedness* serves as a fundamental pairwise comparison from which a relationship can be defined to all words. This forms a basis for the semantic relationships of *similarity*, *paraphrasing* and *analogies*. Error terms arise in the latter higher order relationships due to statistical relationships within the word sets. These errors can be interpreted geometrically with respect to the surface $\mathcal{S}$ on which all PMI vectors lie and, in principle, can be evaluated from higher order statistics (e.g trigrams co-occurrences).

There remain a few further aspects of W2V and Glove to address, e.g. that PMI is weighted over the context window [31], unigram distributions are often raised to the power $^{3}\!/_{4}$ [26], and the probability weighting of Glove [29]. Several open question also arise from our work, in particular regarding the weight vector $\mathbf{v}$ in embedding differences (S.5.2). We hope to address these in future work.

Acknowledgments

We thank Jonathan Mallinson and Iain Murray for helpful comments on this manuscript. Carl Allen and Ivana Balažević were supported by the Centre for Doctoral Training in Data Science, funded by EPSRC (grant EP/L016427/1) and the University of Edinburgh.

References

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *North American Chapter of the Association for Computational Linguistics*, 2009.

[2] Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, 2019.

[3] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 2016.

[4] Fatemeh Torabi Asr, Robert Zinkov, and Michael Jones. Querying word embeddings for similarity and relatedness. In *North American Chapter of the Association for Computational Linguistics*, 2018.

[5] John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526, 2007.

[6] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, 2002.

[7] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Empirical Methods in Natural Language Processing*, 2017.

[8] Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. Explaining and generalizing skip-gram through exponential family principal component analysis. In *European Chapter of the Association for Computational Linguistics*, 2017.

[9] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*, 2018.

[10] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *International Conference on World Wide Web*, 2001.

[11] Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. Skip-Gram - Zipf + Uniform = Vector Additivity. In *Association for Computational Linguistics*, 2017.

[12] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *International Conference on Knowledge Discovery and Data mining*, 2016.

[13] Tatsunori B Hashimoto, David Alvarez-Melis, and Tommi S Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 2016.

[14] Aminul Islam, Evangelos Milios, and Vlado Keselj. Comparing word relatedness measures based on google n-grams. *COLING 2012*, 2012.

[15] Ian Jolliffe. *Principal Component Analysis*. Springer, 2011.

[16] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, 2015.

[17] Andrew J Landgraf and Jeremy Bellay. word2vec Skip-Gram with Negative Sampling is a Weighted Logistic PCA. *arXiv preprint arXiv:1705.09755*, 2017.

[18] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Computational Natural Language Learning*, 2014.

[19] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2014.

[20] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 2015.

[21] Qiuchi Li, Sagar Uprety, Benyou Wang, and Dawei Song. Quantum-inspired complex word embedding. In *Workshop on Representation Learning for NLP*, 2018.

[22] Tal Linzen. Issues in evaluating semantic spaces using word analogies. In *1st Workshop on Evaluating Vector-Space Representations for NLP*, 2016.

[23] Haiming Liu. Quantum-like generalization of complex word embedding: a lightweight approach for textual classification.

[24] Matt Mahoney. text8 wikipedia dump. `http://mattmahoney.net/dc/textdata.html`, 2011. [Online; accessed May 2019].

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[27] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics*, 2013.

[28] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Empirical Methods in Natural Language Processing*, 2017.

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, 2014.

[30] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *International Conference on Knowledge Discovery and Data mining*, 2014.

[31] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying DeepWalk, line, pte, and node2vec. In *International Conference on Web Search and Data Mining*, 2018.

[32] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Workshop on New Challenges for NLP Frameworks*, 2010.

[33] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Empirical Methods in Natural Language Processing*, 2015.

[34] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, 1997.

[35] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

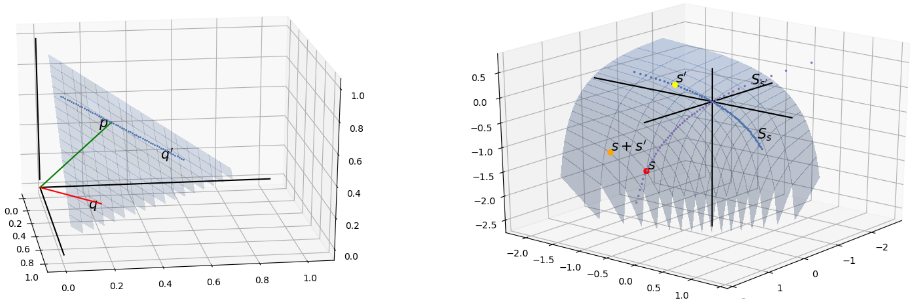
[36] Peter D Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning*. Springer, 2001.

[37] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[38] John Wieting and Douwe Kiela. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations*, 2019.

A Properties of the PMI surface: proofs (Sec 4.1)

1. **$\mathcal{S}$, and any subsurface of $\mathcal{S}$, is non-linear.** This follows directly from the construction of $\mathcal{S}$, in particualr the application of the natural logarithm to the linear surface $\mathcal{R}$.
2. **$\mathcal{S}$ contains the origin $\mathbf{0} \in \mathbb{R}^n$** Follows from construction: $\mathbf{p} = p(\mathcal{E}) \in \mathcal{Q}$ implies $\mathbf{1} = \frac{\mathbf{p}}{p(\mathcal{E})} \in \mathcal{R}$, and threfore $\mathbf{0} == \log \mathbf{1} \in \mathcal{S}$
3. **Probability vector $\mathbf{q}$ is normal to the tangent plane of $\mathcal{S}$** at $\mathbf{s} = \log \mathbf{q}/p(\mathcal{E}) \in \mathcal{S}$. Consider $\mathbf{q} = (q_1, ..., q_n) \in \mathcal{Q}$ as having free parameters $q_{j<n}$ that determine q_n, and let $\mathbf{J} \in \mathbb{R}^{n \times (n-1)}$ define the tangent plane to $\mathcal{S}$ at $\mathbf{s}$, whereby $\mathbf{J}_{i,j} = \frac{\partial s_i}{\partial q_j}$. It can be seen that for $i < n$, $\mathbf{J}_{i,j} = q_j^{-1}$ if $i = j$, and $\mathbf{J}_{i,j} = 0$ otherwise; and tha t$\mathbf{J}_{n,j} = -(\sum_{j=1}^{n-1} q_j)^{-1}$ $\forall j$. It follows that $\mathbf{q}^\top \mathbf{J} = \mathbf{0}$ and $\mathbf{q}$ is therefore normal to the tangent plane.
4. **$\mathcal{S}$ does not intersect with the fully positive or fully negative orthants** (excluding $\mathbf{0}$). This follows from the fact that components of one probability distribution, e.g. $p(\mathcal{E}|w_i)$, cannot *all* be greater (or *all* less) than their counterpart in another, e.g. $p(\mathcal{E})$. Any point in the fully positive or fully negative orthants would contradict this.
5. **The sum of 2 points $\mathbf{s} + \mathbf{s}'$ lies in $\mathcal{S}$ only for certain $\mathbf{s}$, $\mathbf{s}' \in \mathcal{S}$.** For probability vectors $\mathbf{p}$, $\mathbf{q}$, $\mathbf{q}' \in \mathcal{Q}$ and $\mathbf{s} = \log(\mathbf{q}/\mathbf{p})$, $\mathbf{s}' = \log(\mathbf{q}'/\mathbf{p}) \in \mathcal{S}$, we consider operations element-wise with correspondign vector elements denoted by lower case: $\mathbf{s} + \mathbf{s}' \in \mathcal{S}$ *iff* $s + s' = \log(q^*/p)$ for some probability vector $\mathbf{q}^* \in \mathcal{Q}$. Thus, $(q/p)(q'/p) = q^*/p$, or simply $(q/p)q' = q^*$, whereby components $(q/p)q'$ must sum to 1, or in vector terms $(\mathbf{q}/\mathbf{p})^\top \mathbf{q}' = 1$. since $\mathbf{q}'$ is a probability we can also say $(\mathbf{q}/\mathbf{p} - \mathbf{1})^\top \mathbf{q}' = 0$, and we have that $\mathbf{s} + \mathbf{s}' \in \mathcal{S}$ only if $\mathbf{s}' = \log(\mathbf{q}'/\mathbf{p}) \in \mathcal{S}$ with $\mathbf{q}'$ a probability vector orthogonal to $(\mathbf{q}/\mathbf{p}) - \mathbf{1}$. We see that the intersection of the hyperplane orthogonal to $(\mathbf{q}/\mathbf{p}) - \mathbf{1}$ and the simplex defines points $\mathbf{q}'$ that correspond to points in $\mathbf{s}' \in \mathcal{S}$ that can be added to $\mathbf{s}$, i.e. $\mathcal{S}_s$ (See Figs 3a and 3b). Trivially $\mathbf{0} \in \mathcal{S}_s$, $\forall \mathbf{s} \in \mathcal{S}$.



(a) Given point $\mathbf{s}' = \log \mathbf{q}'/\mathbf{p} \in \mathcal{S}$, those $\mathbf{q}'$ on simplex $\mathcal{Q}$ such that $\mathbf{s}' = \log \mathbf{q}'/\mathbf{p}$ satisfies $\mathbf{s} + \mathbf{s}' \in \mathcal{S}$.

(b) Subsurfaces $\mathcal{S}_s$ for a given point $\mathbf{s} \in \mathcal{S}$, and $\mathcal{S}_{s'}$ for any point $\mathbf{s}' \in \mathcal{S}_s$; showing also $\mathbf{s} + \mathbf{s}' \in \mathcal{S}$

Figure 3: Explaining the PMI surface $\mathcal{S}$.

B The *W2V shift*

The number of negative samples per positive sample (k) arises in the optimum of the loss function (4) as the so-called *shift*, $-\log k$, of magnitude comparable to empirical PMI values. It causes dot product interactions to take more negative values, distorting the embeddings. Under certain interactions, aggregate shift terms are seen to cancel [2], but elsewhere have been shown to have a detrimental impact on downstream task performance that removing the *shift* corrects [28]. We show that an *unshifted* relationship between embeddings and PMI exhibit desired properties of word embeddings, whereas the *shift* is an artefact of the *W2V* algorithm chosen as a hyperparameter. We thus drop the *shift* term.

C Further Geometric properties of the PMI surface

Combining both geometric and probabilistic arguments shows:

1. PMI vectors of words w_j that are both conditionally and marginally independent of word w_i, lie in a strict subsurface $\mathcal{S}_{\mathbf{PMI}^i} \subset \mathcal{S}$;
2. only those $\mathbf{PMI}^j \in \mathcal{S}_{\mathbf{PMI}^i}$ add to $\mathbf{PMI}^i$ to give another point on the surface, specifically $\mathbf{PMI}^i + \mathbf{PMI}^j = \mathbf{PMI}^{\{w,w'\}}$ corresponding to the joint occurrence of w_i *and* w_j; and
3. any other $\mathbf{PMI}^{j'} \notin \mathcal{S}_{\mathbf{PMI}^i}$ is off the surface and relates to $\mathbf{PMI}^{\{w,w'\}}$ by an error vector $\boldsymbol{\epsilon}_{i,j}$, reflecting statistical dependence between w_i and w_j.

By symmetry of addition, $\mathbf{PMI}^i \in \mathcal{S}_{\mathbf{PMI}^j}$ *iff* $\mathbf{PMI}^j \in \mathcal{S}_{\mathbf{PMI}^i}$ and such subsurfaces form pairs partitioning $\mathcal{S}$, from which the sum of any pair is in $\mathcal{S}$ and every point in $\mathcal{S}$ is the sum of a unique point pair i.e. $\mathcal{S}_{\mathbf{PMI}^i} \oplus \mathcal{S}_{\mathbf{PMI}^j} = \mathcal{S}$, analogous to the Cartesian product.

D Comparison to embedding relationships of previous works

The following relationships between W2V embeddings and probabilities are assumed in [3]:

$$\mathbf{w}_i = \mathbf{c}_i, \quad \log p(w_i) \approx \tfrac{\|\mathbf{w}_i\|}{2d}^2 - \log Z \quad \text{and} \quad \log p(w_i, c_j) \approx \tfrac{\|\mathbf{w}_i + \mathbf{w}_j\|}{2d}^2 - 2\log Z,$$

By rearranging $\mathbf{w}_i^\top \mathbf{c}_j \approx \text{PMI}(w_i, c_j)$ from (5), it can be shown (proof below) that:

$$\log p(w_i) \approx \tfrac{-\mathbf{w}_i^\top \mathbf{c}_i}{2} + \tfrac{\log p(w_i, c_i)}{2} \quad \text{and} \quad \log p(w_i, c_j) \approx \tfrac{-(\mathbf{w}_i - \mathbf{w}_j)^\top (\mathbf{c}_i - \mathbf{c}_j)}{2} + \tfrac{\log p(w_i, c_i) p(w_j, c_j)}{2}.$$

Having previously seen that $\mathbf{w}_i \neq \mathbf{c}_i$ (Sec 5.1), if we nevertheless make that substitution here to draw comparison, we see that other assumptions differ fundamentally, e.g. having opposite sign, to those relationships directly implied by (5), which is claimed to follow from them. Also, the assumed constant Z can be seen to vary arbitrarily with the extent to which each word co-occurs with itself.

D.1 Proofs

Noting that $p(w_i) = p(c_i)$ since the difference is only the role we attribute to a word, (5) gives:

$$\mathbf{w}_i^\top \mathbf{c}_j \approx \log \tfrac{p(w_i, c_j)}{p(w_i) p(c_j)} = \log p(w_i, c_j) - \log p(w_i) - \log p(w_j) \tag{12}$$

If $i = j$, i.e. target and context words are the same, it follows that:

$$\begin{aligned} \mathbf{w}_i^\top \mathbf{c}_i &\approx \log p(w_i, c_i) - 2\log p(w_i) \\ \text{i.e.} \quad \log p(w_i) &\approx \tfrac{-\mathbf{w}_i^\top \mathbf{c}_i}{2} + \tfrac{\log p(w_i, c_i)}{2} \end{aligned} \tag{13}$$

In the general case:

$$\begin{aligned} (\mathbf{w}_i - \mathbf{w}_j)^\top (\mathbf{c}_i - \mathbf{c}_j) \;&=\; \mathbf{w}_i^\top \mathbf{c}_i - \mathbf{w}_j^\top \mathbf{c}_i - \mathbf{w}_i^\top \mathbf{c}_j + \mathbf{w}_j^\top \mathbf{c}_j \\ &\overset{*}{=}\; \mathbf{w}_i^\top \mathbf{c}_i + \mathbf{w}_j^\top \mathbf{c}_j, -2\mathbf{w}_i^\top \mathbf{c}_j \\ &\overset{(12,13)}{\approx}\; (\log p(w_i, c_i) - 2\log p(w_i)) + (\log p(w_j, c_j) - 2\log p(w_j)) \\ &\qquad - 2(\log p(w_i, c_j) - \log p(w_i) - \log p(w_j)) \\ &=\; \log p(w_i, c_i) + \log p(w_j, c_j) - 2\log p(w_i, c_j) \\ \text{thus} \qquad \log p(w_i, c_j) \;&\approx\; \tfrac{-(\mathbf{w}_i - \mathbf{w}_j)^\top (\mathbf{c}_i - \mathbf{c}_j)}{2} + \tfrac{\log p(w_i, c_i) p(w_j, c_j)}{2}. \end{aligned} \tag{14}$$

The step marked by * relies on $\mathbf{w}_i^\top \mathbf{c}_j = \mathbf{w}_i^\top (\mathbf{I}' \mathbf{w}_j) = (\mathbf{w}_i^\top \mathbf{I}') \mathbf{w}_j = \mathbf{c}_i^\top \mathbf{w}_j = \mathbf{w}_j^\top \mathbf{c}_i$, which follows from the relationship $\mathbf{C} = \mathbf{I}' \mathbf{W}$.

E Why order matters in analogies

Here we look to interpret the finding of Linzen [22], that some words within an analogy are more accurately predicted than others (see their "Reverse (add)") in terms of the explanation of [2], which fits within our overall framework.

From [2], we see that for analogy *"w_a is to w_{a^*} as w_b is to w_{b^*}"*, a "total error" term arises in the relationship $\mathbf{PMI}^{b^*}+\mathbf{PMI}^{a} = \mathbf{PMI}^{a^*}+\mathbf{PMI}^{b}$ between PMI vectors (and thus also word embeddings) due to statistical interactions between word pairs $\{w_a, w_{b^*}\}$ and $\{w_b, w_{a^*}\}$, irrespective of the word considered "missing". That is, when PMI vectors are combined, e.g. $\mathbf{PMI}^{a^*}+\mathbf{PMI}^{b}-\mathbf{PMI}^{a}$ seeking to approximate $\mathbf{PMI}^{b^*}$, the offset to $\mathbf{PMI}^{b^*}$ is given by the total error term (and likewise if we sought $\mathbf{PMI}^{b}$). Although the error term is constant, PMI vectors are not evenly distributed (given words follow a highly non-uniform Zipf distribution), therefore the PMI vector of some words will be in more "cluttered" regions than others (for word embeddings this may be far more the case due to projection to a far lower dimensionality). For such words, say w_{b^*}, the same error term is more likely to allow other PMI vectors to be found closer to the estimated location, given by the linear combination of PMI vectors, than the true PMI vector being sought. In such case it may be possible to solve the analogy to find w_b but not w_{b^*}, even though the error term is the same for both. We note that intuition for this is provided by [22], but we are able to frame this more concretely based on [2].

F Experimental details

F.1 Training

PMI values are pre-computed from the corpus similarly to [29], substituting −1 for missing PMI values. We use the *text8* data set [24] containing $c.$17m tokens and $c.$0.5m unique words (sourced from the English Wikipedia dump, 03/03/06). 5 random word pairs (negative samples) are generated for each true word co-occurrence (positive sample) according to unigram word distributions. Dimensionality is 500. Words appearing less than 5 times are filtered and down-sampling is applied (see [26]). All models converged within 100 epochs (full passes over the PMI matrix). Learning rates that worked well were selected for each model: 0.01 for the least squares models, 0.007 for the W2V loss function. Results are averaged over 3 random seeds.

F.2 Testing

Embeddings are evaluated on relatedness, similarity and analogy tasks using *WordSim353* [10, 1]. Ranking is by cosine similarity and evaluation compares Spearman's correlation between rankings and human-assigned similarity scores. Analogies use Google's analogy data set [25] of $c.$ 20k semantic and syntactic analogy questions 'a is to b as c is to ..?'. Out-of-vocabulary words are filtered as standard [20]. Accuracy is computed by comparing $\operatorname{argmin}_d \|v_a - v_b - v_c + v_d\|$ to the labelled answer.