Analogies Explained: Towards Understanding Word Embeddings

Carl Allen [1] Timothy Hospedales [1]

Abstract

Word embeddings generated by neural network methods such as *word2vec* (W2V) are well known to exhibit seemingly linear behaviour, e.g. the embeddings of analogy *"woman is to queen as man is to king"* approximately describe a parallelogram. This property is particularly intriguing since the embeddings are not trained to achieve it. Several explanations have been proposed, but each introduces assumptions that do not hold in practice. We derive a probabilistically grounded definition of *paraphrasing* that we re-interpret as *word transformation*, a mathematical description of *"w_x is to w_y"*. From these concepts we prove existence of linear relationships between W2V-type embeddings that underlie the analogical phenomenon, identifying explicit error terms.

1. Introduction

The vector representation, or *embedding*, of words underpins much of modern machine learning for natural language processing (e.g. Turney & Pantel (2010)). Where, previously, embeddings were generated explicitly from word statistics, neural network methods are now commonly used to generate *neural embeddings* that are of low dimension relative to the number of words represented, yet achieve impressive performance on downstream tasks (e.g. Turian et al. (2010); Socher et al. (2013)). Of these, *word2vec*[2] (W2V) (Mikolov et al., 2013a) and *Glove* (Pennington et al., 2014) are amongst the best known and on which we focus.

Interestingly, such embeddings exhibit seemingly linear behaviour (Mikolov et al., 2013b; Levy & Goldberg, 2014a), e.g. the respective embeddings of *analogies*, or word relationships of the form *"w_a is to w_{a^*} as w_b is to w_{b^*}"*, often satisfy $\mathbf{w}_{a^*} - \mathbf{w}_a + \mathbf{w}_b \approx \mathbf{w}_{b^*}$, where $\mathbf{w}_i$ is the embedding of word w_i. This enables analogical questions such as "*man is to king as woman is to ..?*" to be solved by vector addition and subtraction. Such high order structure is surprising since word embeddings are trained using only pairwise word co-occurrence data extracted from a text corpus.

We first show that where embeddings factorise *pointwise mutual information* (PMI), it is *paraphrasing* that determines when a linear combination of embeddings equates to that of another word. We say $king$ paraphrases man and $royal$, for example, if there is a semantic equivalence between $king$ and $\{man, royal\}$ combined. We can measure such equivalence with respect to probability distributions over nearby words, in line with Firth's maxim "*You shall know a word by the company it keeps*" (Firth, 1957). We then show that paraphrasing can be reinterpreted as *word transformation* with additive *parameters* (e.g. from man to $king$ by adding $royal$) and generalise to also allow subtraction. Finally, we prove that by interpreting an analogy *"w_a is to w_{a^*} as w_b is to w_{b^*}"* as word transformations w_a to w_{a^*} and w_b to w_{b^*} sharing the same parameters, the linear relationship observed between word embeddings of analogies follows (see overview in Fig 4). Our key contributions are:

- to derive a probabilistic definition of *paraphrasing* and show that it governs the relationship between one (PMI-derived) word embedding and any sum of others;
- to show how paraphrasing can be generalised and interpreted as the *transformation* from one word to another, giving a mathematical formulation for *"w_x is to w_{x^*}"*;
- to provide the first rigorous proof of the linear relationship between word embeddings of analogies, including explicit, interpretable error terms; and
- to show how these relationships materialise between vectors of PMI values, and so too in word embeddings that factorise the PMI matrix, or approximate such a factorisation e.g. W2V and *Glove*.

2. Previous Work

Intuition for the presence of linear analogical relationships, or *linguistic regularity*, amongst word embeddings was first suggested by Mikolov et al. (2013a;b) and Pennington et al. (2014), and has been widely discussed since (e.g. Levy & Goldberg (2014a); Linzen (2016)). More recently, several theoretical explanations have been proposed:

[1] School of Informatics, University of Edinburgh. Correspondence to: Carl Allen <carl.allen@ed.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[2] Throughout, we refer to the more commonly used *Skipgram* implementation of W2V with negative sampling (SGNS).

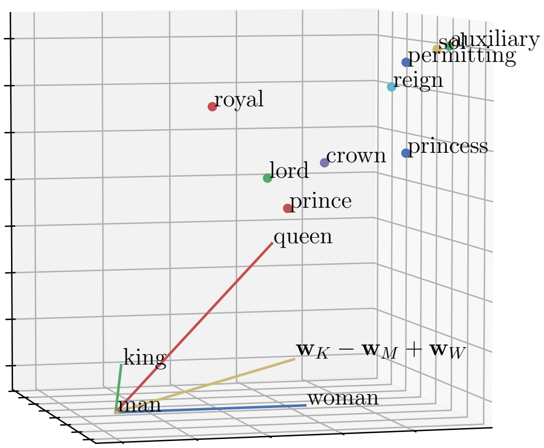


Figure 1. The relative locations of word embeddings for the analogy "*man* is to *king* as *woman* is to ..?". The closest embedding to the linear combination $\mathbf{w}_K - \mathbf{w}_M + \mathbf{w}_W$ is that of *queen*. We explain why this occurs and interpret the difference between them.

- Arora et al. (2016) propose a latent variable model for language that contains several strong *a priori* assumptions about the spatial distribution of word vectors, discussed by Gittens et al. (2017), that we do not require. Also, the two embedding matrices of W2V are assumed equal, which we show to be false in practice.
- Gittens et al. (2017) refer to *paraphrasing*, from which we draw inspiration, but make several assumptions that fail in practice: (i) that words follow a uniform distribution rather than the (highly non-uniform) Zipf distribution; (ii) that W2V learns a conditional distribution – violated by negative sampling (Levy & Goldberg, 2014b); and (iii) that joint probabilities beyond pairwise co-occurrences are zero.
- Ethayarajh et al. (2018) offer a recent explanation based on *co-occurrence shifted PMI*, however that property lacks motivation and several assumptions fail, e.g. it requires more than for opposite sides to have equal length to define a parallelogram in $\mathbb{R}^d$, $d > 2$ (their Lemma 1).

To our knowledge, no previous work mathematically interprets analogies so as to rigorously explain why if *"w_a is to w_{a^*} as w_b is to w_{b^*}"* then a linear relationship manifests between correponding word embeddings.

3. Background

The **Word2Vec** algorithm considers a set of word pairs $\{(w_{i_k}, c_{j_k})\}_k$ generated from a (typically large) text corpus, by allowing the *target* word w_i to range over the corpus, and the *context* word c_j to range over a context window (of size l) symmetric about the target word. For each observed word pair (*positive sample*), k random word pairs (*negative samples*) are generated according to monogram distributions. The 2-layer "neural network" architecture simply multiplies two weight matrices $\mathbf{W}, \mathbf{C} \in \mathbb{R}^{d \times n}$, subject to a non-linear (sigmoid) function, where d is the embedding dimensionality and n is the size of $\mathcal{E}$ the dictionary of unique words in the corpus. Conventionally, $\mathbf{W}$ denotes the matrix closest to the input target words. Columns of $\mathbf{W}$ and $\mathbf{C}$ are the *embeddings* of words in $\mathcal{E}$: $\mathbf{w}_i \in \mathbb{R}^d$ (i^{th} column of $\mathbf{W}$) corresponds to w_i the i^{th} word in $\mathcal{E}$ observed as a target word; and $\mathbf{c}_i \in \mathbb{R}^d$ (i^{th} column of $\mathbf{C}$) corresponds to c_i, the same word when observed as a context word.

Levy & Goldberg (2014b) identified that the objective function for W2V is optimised if:

$$\mathbf{w}_i^\top \mathbf{c}_j \;=\; \mathrm{PMI}(w_i, c_j) - \log k \;, \tag{1}$$

where $\mathrm{PMI}(w_i, c_j) = \log \frac{p(w_i, c_j)}{p(w_i)p(c_j)}$ is known as *pointwise mutual information*. In matrix form, this equates to:

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

where $\mathbf{SPMI}_{i,j} = \mathrm{PMI}(w_i, c_j) - \log k$, (*shifted* PMI).

Glove (Pennington et al., 2014) has the same architecture as W2V. Its embeddings perform comparably and also exhibit linear analogical structure. *Glove*'s loss function is optimised 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. (3) generalises (1) due to the biases, giving *Glove* greater flexibility than W2V and a potentially wider range of solutions. However, we will show that it is factorisation of the PMI matrix that causes linear analogical structure in embeddings, as approximately achieved by W2V (1). We conjecture that the same rationale underpins analogical structure in *Glove* embeddings, perhaps more weakly due to its increased flexibility.

4. Preliminaries

We consider pertinent aspects of the relationship between word embeddings and co-occurrence statistics (1, 2) relevant to the linear structure between embeddings of analogies:

Impact of the *Shift* As a chosen hyper-parameter, reflecting nothing of word properties, any effect on embeddings of k appearing in (1) is arbitrary. Comparing typical values of k with empirical PMI values (Fig 2), shows that the so-called *shift* ($-\log k$) may also be material. Further, it is observed that adjusting the W2V algorithm to avoid any direct impact of the *shift* improves embedding performance (Le, 2017). We conclude that the *shift* is a detrimental artefact of the W2V algorithm and, unless stated otherwise, consider embeddings that factorise the *unshifted* PMI matrix:

$$\mathbf{w}_i^\top \mathbf{c}_j \;=\; \mathrm{PMI}(w_i, c_j) \quad \text{or} \quad \mathbf{W}^\top \mathbf{C} \;=\; \mathbf{PMI} \,. \tag{4}$$

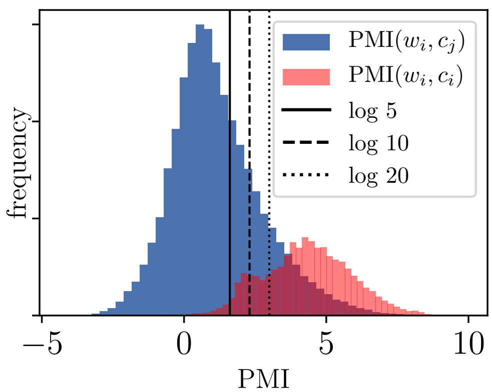


Figure 2. Histogram of PMI(w_i, c_j) for word pairs randomly sampled from text (blue) with PMI(w_i, c_i) for the *same word* overlaid (red, scale enlarged). The *shift* is material for typical values of k.

Reconstruction Error In practice, (2) and (4) hold only *approximately* since $\mathbf{W}^\top\mathbf{C} \in \mathbb{R}^{n\times n}$ is rank-constrained (rank $r \ll d < n$) relative to the factored matrix $\mathbf{M}$, e.g. $\mathbf{M} = \mathbf{PMI}$ in (4). Recovering elements of $\mathbf{M}$ from $\mathbf{W}$ and $\mathbf{C}$ is thus subject to *reconstruction error*. However, we rely throughout on linear relationships in $\mathbb{R}^n$, requiring only that they are sufficiently maintained when projected "down" into $\mathbb{R}^d$, the space of embeddings. To ensure this, we assume:

A1. $\mathbf{C}$ *has full row rank.*

A2. *Letting* $\mathbf{M}_k$ *denote the* k^{th} *column of factored matrix* $\mathbf{M} \in \mathbb{R}^{n\times n}$*, the projection* $f : \mathbb{R}^n \to \mathbb{R}^d$*,* $f(\mathbf{M}_i) = \mathbf{w}_i$ *is approximately homomorphic with respect to addition, i.e.* $f(\mathbf{M}_i + \mathbf{M}_j) \approx f(\mathbf{M}_i) + f(\mathbf{M}_j)$.

A1 is reasonable since $d \ll n$ and d is chosen. A2 means that, whatever the factorisation method used (e.g. analytic, W2V, *Glove*, weighted matrix factorisation (Srebro & Jaakkola, 2003)), linear relationships between columns of $\mathbf{M}$ are sufficiently preserved by columns of $\mathbf{W}$, i.e. the embeddings $\mathbf{w}_i$. For example, minimising a least squares loss function gives the linear projection $\mathbf{w}_i = f_{LSQ}(\mathbf{M}_i) = \mathbf{C}^\dagger\mathbf{M}_i$ for which A2 holds exactly (where $\mathbf{C}^\dagger = (\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C}$, the *Moore-Penrose pseudo-inverse* of $\mathbf{C}^\top$, which exists by A1);[1] whereas for W2V, $\mathbf{w}_i = f_{W2V}(\mathbf{M}_i)$ is non-linear.[2]

Zero Co-occurrence Counts The co-occurrence of rare words are often unobserved, thus their empirical probability estimates zero and PMI estimates undefined. However, for a fixed dictionary $\mathcal{E}$, such zero counts decline as the corpus or context window size increase (the latter can be arbitrarily large if more distant words are down-weighted, e.g. Pennington et al. (2014)). Here, we consider small word sets $\mathcal{W}$ and assume the corpus and context window to be of sufficient size that the *true* values of considered probabilities are non-zero and their PMI values well-defined, i.e.:

A3. $p(\mathcal{W}) > 0, \ \ \forall \mathcal{W} \subseteq \mathcal{E}, |\mathcal{W}| < l,$

where (throughout) "$|\mathcal{W}| < l$" means $|\mathcal{W}|$ *sufficiently less* than l.

The Relationship between W and C Several works (e.g. Hashimoto et al. (2016); Arora et al. (2016)) assume embedding matrices $\mathbf{W}$ and $\mathbf{C}$ to be equal, i.e. $\mathbf{w}_i = \mathbf{c}_i \ \ \forall i$. The assumption is convenient as the number of parameters is halved, equations simplify and consideration of how to use $\mathbf{w}_i$ and $\mathbf{c}_i$ falls away. However, this implies $\mathbf{W}^\top\mathbf{W} = \mathbf{PMI}$, requiring $\mathbf{PMI}$ to be positive semi-definite, which is not true for typical corpora. Thus $\mathbf{w}_i$, $\mathbf{c}_i$ are not equal and modifying W2V to enforce them to be would unnecessarily constrain and may well worsen the low-rank approximation.

5. Paraphrases

Following a similar approach to Gittens et al. (2017), we consider a small set of target words $\mathcal{W} = \{w_1, \dots, w_m\} \subseteq \mathcal{E}$, $|\mathcal{W}| < l$; and the sum of their embeddings $\mathbf{w}_\mathcal{W} = \sum_i \mathbf{w}_i$. In practice, we say word $w_* \in \mathcal{E}$ *paraphrases* $\mathcal{W}$ if w_* and $\mathcal{W}$ are semantically interchangeable within the text, i.e. in circumstances where *all* $w_i \in \mathcal{W}$ appear, w_* could appear instead. This suggests a relationship between the probability distributions $p(c_j|\mathcal{W})$ and $p(c_j|w_*)$, $\forall c_j \in \mathcal{E}$. We refer to such conditional distributions over all context words as the *distribution induced* by $\mathcal{W}$ or w_*, respectively.

5.1. Defining a Paraphrase

Let $\mathcal{C}_\mathcal{W} = \{c_{j_1}, \dots, c_{j_t}\}$ be a sequence of words (with repetition) observed in the context of $\mathcal{W}$.[3] A *paraphrase word* $w_* \in \mathcal{E}$ can be thought of as that which *best explains* the observation of $\mathcal{C}_\mathcal{W}$. From a maximum likelihood perspective we have $w_*^{(1)} = \mathrm{argmax}_{w_i \in \mathcal{E}}\, p(\mathcal{C}_\mathcal{W}|w_i)$. Assuming $c_j \in \mathcal{C}_\mathcal{W}$ to be independent draws from $p(c_j|\mathcal{W})$, gives:

$$\begin{aligned} w_*^{(1)} &= \underset{w_i}{\mathrm{argmax}} \prod_{c_j\in\mathcal{E}} p(c_j|w_i)^{\#_j} \\ &\to \underset{w_i}{\mathrm{argmax}} \sum_{c_j\in\mathcal{E}} p(c_j|\mathcal{W}) \log p(c_j|w_i) \ , \end{aligned}$$

as $|\mathcal{C}_\mathcal{W}| \to \infty$, where $\#_j$ denotes the count of c_j in $\mathcal{C}_\mathcal{W}$. It follows that $w_*^{(1)}$ minimises the Kullback-Leibler (KL) divergence $\Delta_{KL}^{\mathcal{W},w_*}$ between the induced distributions, i.e.:

$$\begin{aligned} \Delta_{KL}^{\mathcal{W},w_*} &= D_{KL}[\, P(c_j|\mathcal{W}) \,||\, P(c_j|w_*) \,] \\ &= \textstyle\sum_j p(c_j|\mathcal{W}) \log \frac{p(c_j|\mathcal{W})}{p(c_j|w_*)} \ . \end{aligned}$$

[1] *w.l.o.g.* we write $f(\cdot) = \mathbf{C}^\dagger(\cdot)$ throughout (except in specific cases) to emphasise linearity of the relationship.

[2] It is beyond the scope of this work to show A2 is satisfied when the W2V loss function is minimised (4). We instead prove existence of linear relationships in the full rank space of PMI columns, thus in linear projections thereof, and assume A2 holds sufficiently for W2V embeddings given (2) and empirical observation of linearity.

[3] By symmetry, $\mathcal{C}_\mathcal{W}$ is the set of target words for which all $w_i \in \mathcal{W}$ are simultaneously observed in the context window.

Alternatively, we might consider $w_*^{(2)}$, the target word whose set of associated context words $\mathcal{C}_{w_*}$ is best explained by $\mathcal{W}$, in the sense that $w_*^{(2)}$ minimises KL divergence $\Delta_{KL}^{w_*,\mathcal{W}} = D_{KL}[P(c_j|w_*)\,||\,P(c_j|\mathcal{W})]$ (where, in general, $\Delta_{KL}^{\mathcal{W},w_*} \neq \Delta_{KL}^{w_*,\mathcal{W}}$). Interpretations of $w_*^{(1)}$ and $w_*^{(2)}$ are discussed in Appendix A. In each case, the KL divergence lower bound (zero) is achieved *iff* the induced distributions are equal, providing a theoretical basis for:

Definition D1. *We say word $w_* \in \mathcal{E}$ paraphrases word set $\mathcal{W} \subseteq \mathcal{E}$, $|\mathcal{W}| < l$, if the paraphrase error $\boldsymbol{\rho}^{\mathcal{W},w_*} \in \mathbb{R}^n$ is (element-wise) small, where:*

$$\boldsymbol{\rho}_j^{\mathcal{W},w_*} \;=\; \log \frac{p(c_j|w_*)}{p(c_j|\mathcal{W})}\;, c_j \in \mathcal{E}.$$

Note that $\mathcal{W}$ and w_* need not appear similarly often for w_* to paraphrase $\mathcal{W}$, only amongst the same context words. We now connect paraphrasing, a semantic relationship, to relationships between word embeddings.

5.2. Paraphrase = Embedding Sum + Error

Lemma 1. *For any word $w_* \in \mathcal{E}$ and word set $\mathcal{W} \subseteq \mathcal{E}$, $|\mathcal{W}| < l$:*

$$\mathrm{PMI}_* = \sum_{w_i \in \mathcal{W}} \mathrm{PMI}_i + \boldsymbol{\rho}^{\mathcal{W},w_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \tau^{\mathcal{W}}\mathbf{1}\;, \qquad (5)$$

where $\mathrm{PMI}_\bullet$ is the column of $\mathbf{PMI}$ corresponding to $w_\bullet \in \mathcal{E}$, $\mathbf{1} \in \mathbb{R}^n$ is a vector of 1s, and error terms $\boldsymbol{\sigma}_j^{\mathcal{W}} = \log \frac{p(\mathcal{W}|c_j)}{\prod_i p(w_i|c_j)}$ and $\tau^{\mathcal{W}} = \log \frac{p(\mathcal{W})}{\prod_i p(w_i)}$.

Proof. (See Appendix B.) As Lem 1 is central to what follows, we sketch its proof: a correspondence is drawn between the product of distributions induced by each $w_i \in \mathcal{W}$ (I) and the distribution induced by w_* (II), by comparison to the distribution induced by joint event $\mathcal{W}$ (III), i.e. observing *all* $w_i \in \mathcal{W}$ in the context window. I relates to III by the (in)dependence of $w_i \in \mathcal{W}$ (i.e. by $\boldsymbol{\sigma}_j^{\mathcal{W}}$, $\tau^{\mathcal{W}}$).[4] II relates to III by the paraphrase error $\boldsymbol{\rho}_j^{\mathcal{W},w_*}$. □

Following immediately from Lem 1 we have:

Theorem 1 (Paraphrase)**.** *For any word $w_* \in \mathcal{E}$ and word set $\mathcal{W} \subseteq \mathcal{E}$, $|\mathcal{W}| < l$:*

$$\mathbf{w}_* = \mathbf{w}_{\mathcal{W}} \;+\; \mathbf{C}^\dagger(\boldsymbol{\rho}^{\mathcal{W},w_*} \;+\; \boldsymbol{\sigma}^{\mathcal{W}} \;-\; \tau^{\mathcal{W}}\mathbf{1})\;, \qquad (6)$$

where $\mathbf{w}_{\mathcal{W}} = \sum_{w_i \in \mathcal{W}} \mathbf{w}_i$.

Proof. Multiply (5) by $\mathbf{C}^\dagger$. □

Thm 1 shows that an embedding (of w_*) and a sum of embeddings (of $\mathcal{W}$) differ by the paraphrase error $\boldsymbol{\rho}^{\mathcal{W},w_*}$ between w_* and $\mathcal{W}$; and $\boldsymbol{\sigma}^{\mathcal{W}}$, $\tau^{\mathcal{W}}$ (collectively *dependence error*) reflecting relationships within $\mathcal{W}$ (unrelated to w_*):

- $\boldsymbol{\sigma}^{\mathcal{W}}$ is a vector reflecting conditional dependencies within $\mathcal{W}$ given each $c_j \in \mathcal{E}$; $\boldsymbol{\sigma}_j^{\mathcal{W}} = 0$ *iff* all $w_i \in \mathcal{W}$ are conditionally independent given each and every $c_j \in \mathcal{E}$;
- $\tau^{\mathcal{W}}$ is a scalar measure of mutual independence of $w_i \in \mathcal{W}$ (thus constant $\forall c_j \in \mathcal{E}$); $\tau^{\mathcal{W}} = 0$ *iff* $w_i \in \mathcal{W}$ are mutually independent.

Corollary 1.1. *A word set $\mathcal{W}$ has no associated dependence error* iff *$w_i \in \mathcal{W}$ are both mutually independent and conditionally independent given each context word $c_j \in \mathcal{E}$.*

Thm 1, which holds for all words w_* and word sets $\mathcal{W}$, explains why and when a paraphrase (e.g. of $\{man, royal\}$ by $king$) can be identified by embedding addition ($\mathbf{w}_{man} + \mathbf{w}_{royal} \approx \mathbf{w}_{king}$). The phenomenon occurs due to a relationship between PMI vectors in $\mathbb{R}^n$ that holds for embeddings in $\mathbb{R}^d$ under projection by $\mathbf{C}^\dagger$ (by A1, A2). The vector error $\mathbf{w}_* - \mathbf{w}_{\mathcal{W}}$ depends on both the paraphrase relationship between w_* and $\mathcal{W}$; and statistical dependencies within $\mathcal{W}$.

Corollary 1.2. *For word $w_* \in \mathcal{E}$ and word set $\mathcal{W} \subseteq \mathcal{E}$, $\mathbf{w}_* \approx \mathbf{w}_{\mathcal{W}}$ if w_* paraphrases $\mathcal{W}$ and $w_i \in \mathcal{W}$ are materially independent (i.e. net dependence error is small).*

5.3. Do Linear Relationships Identify Paraphrases?

The converse of Cor 1.2 is false: $\mathbf{w}_* \approx \mathbf{w}_{\mathcal{W}}$ does not imply w_* paraphrases $\mathcal{W}$. Specifically, *false positives* arise if: (i) paraphrase and dependence error terms are material but happen to cancel, i.e. *total error* $\boldsymbol{\rho}^{\mathcal{W},w_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \tau^{\mathcal{W}}\mathbf{1} \approx \mathbf{0}$; or (ii) material components of the total error fall within the high $(n-d)$ dimensional null space of $\mathbf{C}^\dagger$ and project to a small vector difference between $\mathbf{w}_*$ and $\mathbf{w}_{\mathcal{W}}$. Case (i) can arise in PMI vectors (Lem 1) and thus lower rank embeddings also (Thm 1), but is highly unlikely in practice due to the high dimensionality (n). Case (ii) can arise only in lower rank embeddings (Thm 1) and might be minimised by a good choice of factorisation or projection method.

5.4. Paraphrasing in *Explicit* Embeddings

Lem 1 applies to full rank $\mathbf{PMI}$ vectors, without reconstruction error or case (ii) false positives (Sec 5.3), explaining the linear relationships observed by Levy & Goldberg (2014a).

Corollary 1.3. *Thm 1 holds for* explicit *word embeddings, i.e. columns of* $\mathbf{PMI}$.

Proof. Choose factorisation $\mathbf{W} = \mathbf{PMI}$, $\mathbf{C} = \mathbf{I}$ (the identity matrix) in Thm 1. □

5.5. Paraphrasing in W2V Embeddings

Thm 1 extends to W2V embeddings by substituting $\mathbf{v}_i{}^\top \mathbf{v}'_j = \mathrm{PMI}(w_i, c_j) - \log k$ and f_{W2V}:

Corollary 1.4. *Under conditions of Thm 1, W2V embeddings satisfy:*

$$\mathbf{w}_* = \mathbf{w}_{\mathcal{W}} + f_{W2V}\big(\boldsymbol{\rho}^{\mathcal{W},w_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \tau^{\mathcal{W}}\mathbf{1} + \log k(|\mathcal{W}|-1)\mathbf{1}\big)\,. \qquad (7)$$

[4] Analogous to a product of marginal probabilities relating to their joint probability subject to independence.

Comparing (6) and (7) shows that paraphrases correspond to linear relationships in W2V embeddings with an additional error term linear in $|\mathcal{W}|$, and hence with less accuracy if $|\mathcal{W}|>1$, than for embeddings that factorise $\mathbf{PMI}$.

6. Analogies

An *analogy* is said to hold for words $w_a, w_{a^*}, w_b, w_{b^*} \in \mathcal{E}$ if, in some sense, *"w_a is to w_{a^*} as w_b is to w_{b^*}"*. Since in principle the same relationship may extend further ("... as w_c is to w_{c^*}" etc), we characterise a general analogy $\mathfrak{A}$ by a set of ordered word pairs $S_{\mathfrak{A}} \subseteq \mathcal{E}\times\mathcal{E}$, where $(w_x, w_{x^*}) \in S_{\mathfrak{A}}$, $w_x, w_{x^*} \in \mathcal{E}$, *iff* "$w_x$ is to w_{x^*} as ... [all other analogical pairs]" under $\mathfrak{A}$. Our aim is to explain why respective word embeddings often satisfy:

$$\mathbf{w}_{b^*} \;\approx\; \mathbf{w}_{a^*} - \mathbf{w}_a + \mathbf{w}_b\ , \tag{8}$$

or why in the more general case:

$$\mathbf{w}_{x^*} - \mathbf{w}_x \;\approx\; \mathbf{u}_{\mathfrak{A}}\ , \tag{9}$$

$\forall (w_x, w_{x^*}) \in S_{\mathfrak{A}}$ and vector $\mathbf{u}_{\mathfrak{A}} \in \mathbb{R}^n$ specific to $\mathfrak{A}$.

We split the task of understanding why analogies give rise to Equations 8 and 9 into: **Q1**) understanding conditions under which word embeddings can be added and subtracted to approximate other embeddings; **Q2**) establishing a mathematical interpretation of "w_x is to w_{x^*}"; and **Q3**) drawing a correspondence between those results. We show that all of these can be answered with paraphrasing by generalising the notion to word sets.

6.1. Paraphrasing Word Sets

Definition D2. *We say word set $\mathcal{W}_* \subseteq \mathcal{E}$ <u>paraphrases</u> word set $\mathcal{W} \subseteq \mathcal{E}$, $|\mathcal{W}|, |\mathcal{W}_*| < l$, if <u>paraphrase error</u> $\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} \in \mathbb{R}^n$ is (element-wise) small, where:*

$$\boldsymbol{\rho}_j^{\mathcal{W},\mathcal{W}_*} \;=\; \log \tfrac{p(c_j|\mathcal{W}_*)}{p(c_j|\mathcal{W})}\ , c_j \in \mathcal{E}.$$

D2 generalises D1 such that the paraphrase term $\mathcal{W}_*$, previously w_*, can be more than one word.[5] Analogously to D1, word sets paraphrase one another if they induce equivalent distributions over context words. Note that paraphrasing under D2 is both reflexive and symmetric (since $|\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*}| = |\boldsymbol{\rho}^{\mathcal{W}_*,\mathcal{W}}|$), thus "$\mathcal{W}_*$ paraphrases $\mathcal{W}$" and "$\mathcal{W}$ paraphrases $\mathcal{W}_*$" are equivalent and denoted $\mathcal{W} \approx_{\text{P}} \mathcal{W}_*$.

Analogues of Lem 1 and Thm 1 follow:

Lemma 2. *For any word sets $\mathcal{W}$, $\mathcal{W}_* \subseteq \mathcal{E}$, $|\mathcal{W}|$, $|\mathcal{W}_*| < l$:*

$$\sum_{w_i \in \mathcal{W}_*} \text{PMI}_i = \sum_{w_i \in \mathcal{W}} \text{PMI}_i + \boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}\,. \tag{10}$$

Proof. (See Appendix C.) □

[5]Equivalently, D1 is a special case of D2 with $|\mathcal{W}_*| = 1$, hence we reuse terms without ambiguity.

Theorem 2 (Generalised Paraphrase)**.** *For any word sets $\mathcal{W}$, $\mathcal{W}_* \subseteq \mathcal{E}$, $|\mathcal{W}|, |\mathcal{W}_*| < l$:*

$$\mathbf{w}_{\mathcal{W}_*} = \mathbf{w}_{\mathcal{W}} + \mathbf{C}^{\dagger}(\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1})\ .$$

Proof. Multiply (10) by $\mathbf{C}^{\dagger}$. □

Note that $|\mathcal{W}_*| = 1$ recovers Lem 1 and Thm 1. With analogies in mind, we restate Thm 2 as:

Corollary 2.1. *For any words $w_x, w_{x^*} \in \mathcal{E}$ and word sets $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$, $|\mathcal{W}^+|, |\mathcal{W}^-| < l-1$:*

$$\mathbf{w}_{x^*} = \mathbf{w}_x + \mathbf{w}_{\mathcal{W}^+} - \mathbf{w}_{\mathcal{W}^-} + \mathbf{C}^{\dagger}(\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}), \tag{11}$$

where $\mathcal{W} = \{w_x\} \cup \mathcal{W}^+$, $\mathcal{W}_* = \{w_{x^*}\} \cup \mathcal{W}^-$.

Proof. Set $\mathcal{W} = \{w_x\} \cup \mathcal{W}^+$, $\mathcal{W}_* = \{w_{x^*}\} \cup \mathcal{W}^-$ in Thm 2. □

Cor 2.1 shows how any word embedding $\mathbf{w}_{x^*}$ relates to a linear combination of other embeddings ($\mathbf{w}_{\Sigma} = \mathbf{w}_x + \mathbf{w}_{\mathcal{W}^+} - \mathbf{w}_{\mathcal{W}^-}$), due to an equivalent relationship between columns of $\mathbf{PMI}$. Analogously to one-word (D1) paraphrases, the vector difference $\mathbf{w}_{x^*} - \mathbf{w}_{\Sigma}$ depends on the paraphrase error that reflects the relationship <u>between</u> the two word sets $\mathcal{W}_*$, $\mathcal{W}$; and the dependence error that reflects statistical dependence between words <u>within</u> each of $\mathcal{W}$ and $\mathcal{W}_*$.

Corollary 2.2. *For terms as defined above, $\mathbf{w}_{x^*} \approx \mathbf{w}_x + \mathbf{w}_{\mathcal{W}^+} - \mathbf{w}_{\mathcal{W}^-}$ if $\mathcal{W}_* \approx_{\text{P}} \mathcal{W}$ and $w_i \in \mathcal{W}$ and $w_i \in \mathcal{W}_*$ are materially independent or dependence terms materially cancel.*

False positives can arise as discussed in Sec 5.3.

6.2. From Paraphrases to Analogies

A special case of Cor 2.1 gives:

Corollary 2.3. *For any $w_a, w_{a^*}, w_b, w_{b^*} \in \mathcal{E}$:*

$$\mathbf{w}_{b^*} = \mathbf{w}_{a^*} - \mathbf{w}_a + \mathbf{w}_b + \mathbf{C}^{\dagger}(\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1})\,, \tag{12}$$

where $\mathcal{W} = \{w_b, w_{a^*}\}$ and $\mathcal{W}_* = \{w_{b^*}, w_a\}$.

Proof. Set $w_x = w_b$, $w_{x^*} = w_{b^*}$, $\mathcal{W}^+ = \{w_{a^*}\}$, $\mathcal{W}^- = \{w_a\}$ in Cor 2.1. □

Thus we see that (8) holds *if* $\{w_{b^*}, w_a\} \approx_{\text{P}} \{w_b, w_{a^*}\}$ and those word pairs exhibit *similar dependence* (Sec 6.6). More generally, by Cor 2.1 we see that (9) is satisfied by $\mathbf{u}_{\mathfrak{A}} \approx \mathbf{w}_{\mathcal{W}^+} - \mathbf{w}_{\mathcal{W}^-}$ *if* $\{w_{x^*}, \mathcal{W}^-\} \approx_{\text{P}} \{w_x, \mathcal{W}^+\}$ $\forall (w_x, w_{x^*}) \in S_{\mathfrak{A}}$ for *common* word sets $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$ and each pair of paraphrasing word sets exhibit similar dependence.

This establishes sufficient conditions for the linear relationships observed in analogy embeddings (8, 9) in terms of semantic relationships, answering Q1. However, those relationships are *paraphrases*, with no obvious connection to the "w_x *is to* w_{x^*}..." relationships of analogies. We now show that paraphrases sufficient for (8, 9) correspond to analogies by introducing the concept of *word transformation*.

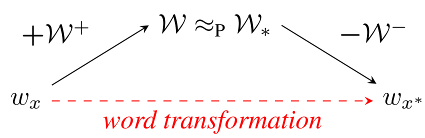
6.3. Word Transformation

The paraphrase of a word set $\mathcal{W}$ by word w_* (D1) has, so far, been considered in terms of an equivalence between $\mathcal{W}$ and w_* by reference to their induced distributions. Alternatively, that paraphrase can be interpreted as a *transformation* from an arbitrary $w_s \in \mathcal{W}$ to w_* by adding words $\mathcal{W}^+ = \{w_i \in \mathcal{W}, w_i \neq w_s\}$. Notionally, $\mathcal{W}^+$ can be considered "words that make w_s more like w_*". More precisely, $w_i \in \mathcal{W}^+$ *add context* to w_s: we move from a distribution induced by w_s alone to one induced by the *joint* event of simultaneously observing w_s and all $w_i \in \mathcal{W}^+$, a *contextualised* occurrence of w_s with an induced distribution closer that of w_*. A similar view can be taken of the associated embedding addition: starting with $\mathbf{w}_s$, add $\mathbf{w}_i\ \forall w_i \in \mathcal{W}^+$ to approximate $\mathbf{w}_*$. Note that only *addition* applies.

Moving to D2, the paraphrase of one word set $\mathcal{W}$ by another $\mathcal{W}_*$ can be interpreted additively as starting with some $w_x \in \mathcal{W}$, $w_{x^*} \in \mathcal{W}_*$, and adding $\mathcal{W}^+ = \{w_i \in \mathcal{W}, w_i \neq w_x\}$, $\mathcal{W}^- = \{w_i \in \mathcal{W}_*, w_i \neq w_{x^*}\}$, respectively, such that the resulting sets $\mathcal{W}$ and $\mathcal{W}_*$ induce similar distributions, i.e. paraphrase. In effect, context is added to both w_x and w_{x^*} until their contextualised cases $\mathcal{W}$ and $\mathcal{W}_*$ paraphrase (Fig 3a). Note $\mathcal{W}$ and $\mathcal{W}_*$ may have no intuitive meaning and need not correspond to a single word, unlike D1 paraphrases. Alternatively, such a paraphrase can be interpreted as a transformation from $w_x \in \mathcal{W}$ to $w_{x^*} \in \mathcal{W}^*$ by adding $w_i \in \mathcal{W}^+$ and *subtracting* $w_i \in \mathcal{W}^-$. "Subtraction" is effected by *adding words to the other side*, i.e. to w_{x^*}.[6] Just as adding words to w_x adds or *narrows* its context, subtracting words removes or *broadens* context. Context is thus added and removed to transform from w_x to w_{x^*}, in which the paraphrase between $\mathcal{W}$ and $\mathcal{W}_*$ effectively serves as an intermediate step (Fig 3b). We refer to $\mathcal{W}^+$, $\mathcal{W}^-$ as *transformation parameters*, which can be thought of as *explaining the difference* between w_x and w_{x^*} with a "richer dictionary" than that available to D1 paraphrases by including *differences* between words. More precisely, transformation parameters align the induced distributions to create a paraphrase.

This interpretation show equivalence between a paraphrase $\mathcal{W} \approx_{\text{P}} \mathcal{W}_*$ and a word transformation – a relationship between $w_x \in \mathcal{W}$ and $w_{x^*} \in \mathcal{W}_*$ based on the addition and subtraction of context that is mirrored in the addition and subtraction of embeddings. Mathematical equivalence of the perspectives is reinforced by an alternate proof of Cor 2.1 in Appendix D that begins with terms in only w_x and w_{x^*}, highlighting that *any* words $\mathcal{W}^+$, $\mathcal{W}^-$ can be introduced, but only certain choices form the necessary paraphrase.

(a) Adding context to each of w_x and w_{x^*} to reach a paraphrase.



(b) Adding and subtracting context to *transform* w_x to w_{x^*}.

Figure 3. Perspectives of the paraphrase $\mathcal{W} \approx_{\text{P}} \mathcal{W}_*$.

Definition D3. *There exists a* <u>word transformation</u> *from* $w_x \in \mathcal{E}$ *to* $w_{x^*} \in \mathcal{E}$ *with* <u>transformation parameters</u> $\mathcal{W}^+$, $\mathcal{W}^- \subseteq \mathcal{E}$ iff $\{w_x\} \cup \mathcal{W}^+ \approx_{\text{P}} \{w_{x^*}\} \cup \mathcal{W}^-$.

Note that transformation parameters may not be unique and always (trivially) include $\mathcal{W}^+ = \{w_{x^*}\}$, $\mathcal{W}^- = \{w_x\}$.

6.4. Interpreting "*a is to a* as b is to b**"

With word transformation as a means of describing semantic difference between words, we mathematically interpret analogies. Specifically, we consider "w_x *is to* w_{x^*}" to refer to a transformation from w_x to w_{x^*} and an analogy to require an equivalence between such word transformations.

Definition D4. *We say* "w_a is to w_{a^*} as w_b is to w_{b^*}" *for* $w_a, w_b, w_{a^*}, w_{b^*} \in \mathcal{E}$ iff *there exist parameters* $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$ *that simultaneously transform* w_a *to* w_{a^*} *and* w_b *to* w_{b^*}.

We show that the linear relationships between word embeddings of analogies (8, 9) follow from D4.

Lemma 3. *If* "w_a is to w_{a^*} as w_b is to w_{b^*}" *by D4 with transformation parameters* $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$, *then:*

$$\begin{aligned}\text{PMI}_{b^*} = \text{PMI}_{a^*} - \text{PMI}_a + \text{PMI}_b& \\ + \boldsymbol{\rho}^{\mathcal{W}^b, \mathcal{W}^b_*} - \boldsymbol{\rho}^{\mathcal{W}^a, \mathcal{W}^a_*}& \\ + (\boldsymbol{\sigma}^{\mathcal{W}^b} - \boldsymbol{\sigma}^{\mathcal{W}^b_*}) - (\boldsymbol{\sigma}^{\mathcal{W}^a} - \boldsymbol{\sigma}^{\mathcal{W}^a_*})& \\ - ((\tau^{\mathcal{W}^b} - \tau^{\mathcal{W}^b_*}) - (\tau^{\mathcal{W}^a} - \tau^{\mathcal{W}^a_*}))\mathbf{1},& \quad (13)\end{aligned}$$

where $\mathcal{W}^x = \{w_x\} \cup \mathcal{W}^+$, $\mathcal{W}^x_* = \{w_{x^*}\} \cup \mathcal{W}^-$ *for* $x \in \{a, b\}$ *and* $\boldsymbol{\rho}^{\mathcal{W}^b, \mathcal{W}^b_*}$, $\boldsymbol{\rho}^{\mathcal{W}^a, \mathcal{W}^a_*}$ *are small.*

Proof. Let $\mathcal{W} = \mathcal{W}^x$, $\mathcal{W}_* = \mathcal{W}^x_*$ for $x \in \{a, b\}$ in instances of Cor 2.1 and take the difference. $\mathcal{W}^x$ paraphrases $\mathcal{W}^x_*$ for $x \in \{a, b\}$ by D3 and D4. □

[6] Analogous to standard algebra: if $x < y$, equality is achieved either by adding to x or by subtracting from y.

$$\begin{array}{c}\text{“}w_a \textit{ is to } w_{a^*}\\ \textit{as}\\ w_b \textit{ is to } w_{b^*}\text{”}\end{array} \iff \begin{array}{c} w_a \xrightarrow[\mathcal{W}^-]{\mathcal{W}^+} w_{a^*} \\ \wedge \\ w_b \xrightarrow[\mathcal{W}^-]{\mathcal{W}^+} w_{b^*}\end{array} \iff \begin{array}{c}\{w_a, \mathcal{W}^+\} \approx_{\mathrm{P}} \{w_{a^*}, \mathcal{W}^-\} \\ \wedge \\ \{w_b, \mathcal{W}^+\} \approx_{\mathrm{P}} \{w_{b^*}, \mathcal{W}^-\}\end{array} \implies \begin{array}{c}\mathbf{w}_{a^*} - \mathbf{w}_a \\ \approx \\ \mathbf{w}_{b^*} - \mathbf{w}_b\end{array}$$

Figure 4. Summary of steps to prove the relationship between analogies and word embeddings (omitting dependence error). $w_x \xrightarrow[\mathcal{W}^-]{\mathcal{W}^+} w_{x^*}$ denotes a word transformation w_x to w_{x^*} with parameters $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$.

Theorem 3 (Analogies). *If* "w_a is to w_{a^*} as w_b is to w_{b^*}" *by D4 with* $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$ *, then:*

$$\begin{aligned}\mathbf{w}_{b^*} = \ & \mathbf{w}_{a^*} - \mathbf{w}_a + \mathbf{w}_b \\ & + \mathbf{C}^\dagger(\boldsymbol{\rho}^{\mathcal{W}^b, \mathcal{W}^b_*} - \boldsymbol{\rho}^{\mathcal{W}^a, \mathcal{W}^a_*} \\ & + (\boldsymbol{\sigma}^{\mathcal{W}^b} - \boldsymbol{\sigma}^{\mathcal{W}^b_*}) - (\boldsymbol{\sigma}^{\mathcal{W}^a} - \boldsymbol{\sigma}^{\mathcal{W}^a_*}) \\ & - ((\tau^{\mathcal{W}^b} - \tau^{\mathcal{W}^b_*}) - (\tau^{\mathcal{W}^a} - \tau^{\mathcal{W}^a_*}))\mathbf{1}).\end{aligned}$$

with terms as defined in Lem 3.

Proof. Multiply (13) by $\mathbf{C}^\dagger$. □

More generally, if D4 applies for a set of ordered word pairs $S = \{(w_x, w_{x^*})\}$, i.e. "w_a *is to* w_{a^*} *as* w_b *is to* w_{b^*}" $\forall (w_a, w_{a^*}), (w_b, w_{b^*}) \in S$ with transformation parameters $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$, then each set $\{w_{x^*}, \mathcal{W}^-\}$ must paraphrase $\{w_x, \mathcal{W}^+\}$ by D3, and (11) holds with small paraphrase error. By this and Thm 3 we know that word embeddings of an analogy $\mathbf{w}_a, \mathbf{w}_b, \mathbf{w}_{a^*}, \mathbf{w}_{b^*}$ satisfy linear relationships (8, 9), subject to dependence error.

A few questions remain: how to find appropriate transformation parameters; and, given non-uniqueness, which to choose? Addressing these in reverse order:

Transformation Parameter Equivalence

By Lem 3, if "w_a *is to* w_{a^*} *as* w_b *is to* w_{b^*}" then, subject to dependence error:

$$\mathrm{PMI}_{b^*} - \mathrm{PMI}_b \approx \mathrm{PMI}_{a^*} - \mathrm{PMI}_a \ . \tag{14}$$

If parameters $\mathcal{W}_2^+, \mathcal{W}_2^-$ exist that (*w.l.o.g.*) transform w_a to w_{a^*} then (13) holds by suitably redefining $\mathcal{W}^x$, $\mathcal{W}^x_*$, in which $\boldsymbol{\rho}^{\mathcal{W}^a, \mathcal{W}^a_*}$ is small but nothing is known of $\boldsymbol{\rho}^{\mathcal{W}^b, \mathcal{W}^b_*}$. Thus, subject to dependence error:

$$\mathrm{PMI}_{b^*} - \mathrm{PMI}_b \approx \mathrm{PMI}_{a^*} - \mathrm{PMI}_a + \boldsymbol{\rho}^{\mathcal{W}^b, \mathcal{W}^b_*} \ . \tag{15}$$

By (14), (15), subject to dependence error, $\boldsymbol{\rho}^{\mathcal{W}^b, \mathcal{W}^b_*}$ is also small and $\mathcal{W}_2^+, \mathcal{W}_2^-$ must also transform w_b to w_{b^*}. Thus transformation parameters of any analogical pair transform all pairs and all applicable transformation parameters can be considered equivalent, up to dependence error.

Corollary 3.1. *For analogy* $\mathfrak{A}$*, if parameters* $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$ *transform* w_x *to* w_{x^*} *for any* $(w_x, w_{x^*}) \in S_{\mathfrak{A}}$*, then* $\mathcal{W}^+$*,* $\mathcal{W}^-$ *simultaneously transform* w_x *to* w_{x^*} $\forall (w_x, w_{x^*}) \in S_{\mathfrak{A}}$*.*

Identifying Transformation Parameters

To identify "words that explain the difference between other words" might, in general, be non-trivial. However, by Cor 3.1, transformation parameters for analogy $\mathfrak{A}$ can simply be chosen as $\mathcal{W}^+ = \{w_{x^*}\}$, $\mathcal{W}^- = \{w_x\}$ for any $(w_x, w_{x^*}) \in S_{\mathfrak{A}}$.[7] Making an arbitrary choice, Thm 3 simplifies to:

Corollary 3.2. *If* "w_a is to w_{a^*} as w_b is to w_{b^*}" *then:*

$$\begin{aligned}\mathbf{w}_{b^*} = \mathbf{w}_{a^*} - \mathbf{w}_a + \mathbf{w}_b + \mathbf{C}^\dagger(&\boldsymbol{\rho}^{\mathcal{W}, \mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} \\ & -(\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}),\end{aligned} \tag{16}$$

where $\mathcal{W} = \{w_b, w_{a^*}\}$*,* $\mathcal{W}_* = \{w_{b^*}, w_a\}$ *and* $\boldsymbol{\rho}^{\mathcal{W}, \mathcal{W}_*}$ *is small.*

Proof. Let $\mathcal{W}^+ = \{w_{a^*}\}, \mathcal{W}^- = \{w_a\}$ in Thm 3. □

We arrive back at (12) but now link directly to analogies, proving that word embeddings of analogies satisfy linear relationships (8) and (9), subject to dependence error. Fig 4 shows a summary of all steps to prove Cor 3.2. D4 also provides a mathematical interpretation of what we mean when we say "w_a *is to* w_{a^*} *as* w_b *is to* w_{b^*}".

6.5. Example

To demonstrate the concepts developed, we consider the canonical analogy $\mathfrak{A}^*$: *"man is to king as woman is to queen"*, for which $S_{\mathfrak{A}^*} = \{(man,\ king), (woman,\ queen)\}$. By D4, there exist parameters $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$ that simultaneously transform man to $king$ and $woman$ to $queen$, which (by Cor 3.1) can be chosen to be $\mathcal{W}^+ = \{queen\}$, $\mathcal{W}^- = \{woman\}$. Thus $\mathfrak{A}^*$ implies that $\{man,\ queen\} \approx_{\mathrm{P}} \{king,\ woman\}$ and $\{woman,\ queen\} \approx_{\mathrm{P}} \{queen,\ woman\}$, the latter being trivially true. By Cor 2.1, $\mathfrak{A}^*$ therefore implies:

$$\begin{aligned}\mathbf{w}_Q \ = \ \mathbf{w}_K - \mathbf{w}_M + \mathbf{w}_W + \mathbf{C}^\dagger(&\boldsymbol{\rho}^{\mathcal{W}, \mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} \\ & - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}) \ ,\end{aligned}$$

where we abbreviate words by their initials and, explicitly:

$$\boldsymbol{\rho}^{\mathcal{W}, \mathcal{W}_*} = \log \frac{p(c_j | w_Q, w_M)}{p(c_j | w_W, w_K)} \qquad \text{(which must be small)},$$

$$\boldsymbol{\sigma}^{\mathcal{W}} = \log \frac{p(w_W, w_K | c_j)}{p(w_W | c_j) p(w_K | c_j)}, \qquad \tau^{\mathcal{W}} = \log \frac{p(w_W, w_K)}{p(w_W) p(w_K)},$$

$$\boldsymbol{\sigma}^{\mathcal{W}_*} = \log \frac{p(w_Q, w_M | c_j)}{p(w_Q | c_j) p(w_M | c_j)}, \qquad \tau^{\mathcal{W}_*} = \log \frac{p(w_Q, w_M)}{p(w_Q) p(w_M)} \ .$$

[7] In the case of an analogical question "w_a *is to* w_{a^*} *as* w_b *is to ... ?*", there is only one choice: $\mathcal{W}^+ = \{w_{a^*}\}$, $\mathcal{W}^- = \{w_a\}$.

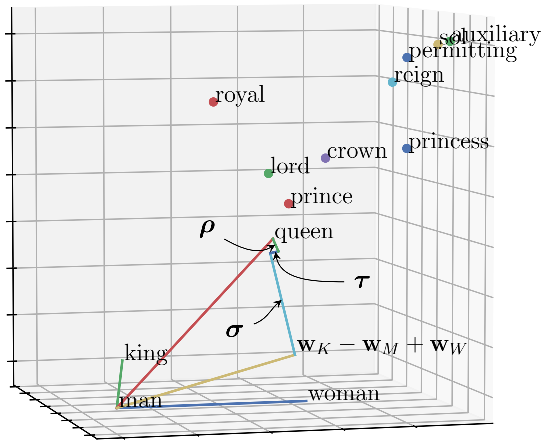


Figure 5. The plot shows the same embeddings of Fig 1, now with the difference between $\mathbf{w}_K - \mathbf{w}_M + \mathbf{w}_W$ and the embedding of $queen$ explained (see connecting "zigzag") as the sum of conditional independence error ($\boldsymbol{\sigma}$), independence error ($\boldsymbol{\tau} = \tau\mathbf{1}$) and paraphrase error ($\boldsymbol{\rho}$). As anticipated, their sum is smallest for $queen$. Related words are seen nearby, with unrelated words clustered further away. Plot generated by fixing the xy plane to contain man, $king$, $queen$ and all other vectors plotted relatively, i.e. the z-axis captures any component off the xy-plane. Values are computed from the "text8" corpus (Mahoney, 2011).

Thus $\mathbf{w}_Q \approx \mathbf{w}_K - \mathbf{w}_M + \mathbf{w}_W$ subject to the accuracy with which $\{man, queen\}$ paraphrases $\{king, woman\}$ and statistical dependencies within those word pairs (see Fig 5).

6.6. Dependence error in analogies

Dependence error terms for analogies (13) bear an important distinction from those in one-word paraphrases (5). When a word set $\mathcal{W}$ is paraphrased by a single word w_*, the dependence error comprises a conditional independence term ($\boldsymbol{\sigma}^{\mathcal{W}}$) and a mutual independence term ($\tau^{\mathcal{W}}\mathbf{1}$) that bear no obvious relationship to one another and can only cancel by chance, which is low in high dimensions. However, (13) contains offsetting pairs of each component ($\boldsymbol{\sigma}^{\mathcal{W}}, \boldsymbol{\sigma}^{\mathcal{W}_*}, \tau^{\mathcal{W}}, \tau^{\mathcal{W}_*}$), i.e. terms of the same form that may cancel, thus word sets with *similar dependence terms* will paraphrase with small overall dependence error.

It is illustrative to consider the case $w_a = w_b$, $w_{a^*} = w_{b^*}$, corresponding to the trivial analogy "w_a is to w_{a^*} as "w_a is to w_{a^*}", which holds true with zero total error for any word pair. Considering specific error terms: the paraphrase error is zero since $p(c_j|\{w_a, w_{a^*}\}) = p(c_j|\{w_{a^*}, w_a\})$, $\forall c_j \in \mathcal{E}$, thus the net dependence error is also zero. However, individual dependence error terms, e.g. $\log \frac{p(w_a, w_{a^*})}{p(w_a)p(w_{a^*})}$, are generally non-zero. This therefore proves existence of a case in which non-zero dependence error terms negate one another to give a negligible net dependence error.

6.7. Analogies in *explicit* embeddings

As with paraphrases, analogical relationships in embeddings stem from relationships between columns of $\mathbf{PMI}$.

Corollary 3.3. *Cor 3.2 applies to explicit (full-rank) embeddings, i.e. columns of* $\mathbf{PMI}$*, with* $\mathbf{C} = \mathbf{I}$ *(the identity matrix).*

6.8. Analogies in W2V embeddings

As with paraphrases (Sec 5.5), the results for analogies can be extended to W2V embeddings by including the *shift* term appropriately throughout. Since the transformation parameters for analogies are of equal size (i.e. $|\mathcal{W}^+| = |\mathcal{W}^-| = 1$), we find that all *shift* terms cancel.

Corollary 3.4. *Cor 3.2 applies to W2V embeddings replacing the projection* $\mathbf{C}^\dagger(\cdot)$ *with* $f_{W2V}(\cdot)$*.*

Thus, linear relationships between embeddings for analogies hold equally for W2V embeddings as for those derived without the *shift* distortion. Whilst perhaps surprising, this is corroborative since linear analogical relationships have been observed extensively in W2V embeddings (e.g. Levy & Goldberg (2014a)), as is now justified theoretically. Thus we know that analogies hold for W2V embeddings subject to higher order statistical relationships between words of the analogy as defined by the paraphrase and dependence errors.

7. Conclusion

In this work, we develop a probabilistically principled definition of *paraphrasing* by which equivalence is drawn between words and word sets by reference to the distributions they induce over words around them. We prove that, subject to statistical dependencies, paraphrase relationships give rise to linear relationships between word embeddings that factorise PMI (including columns of the PMI matrix), and thus others that approximate such a factorisation, e.g. W2V and *Glove*. By showing that paraphrases can be interpreted as *word transformations*, we enable analogies to be mathematically defined and, thereby, properties of semantics to be translated into properties of word embeddings. This provides the first rigorous explanation for the presence of linear relationships between the word embeddings of analogies.

In future work we aim to extend our understanding of the relationships between word embeddings to other applications of discrete object representation that rely on an underlying matrix factorisation, e.g. graph embeddings and recommender systems. Also, word embeddings are known to capture stereotypes present in corpora (Bolukbasi et al. (2016)) and future work may look at developing our understanding of embedding composition to foster principled methods to correct or *debias* embeddings.

Acknowledgements

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

References

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 2016.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 2016.

Ethayarajh, K., Duvenaud, D., and Hirst, G. Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*, 2018.

Firth, J. R. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, 1957.

Gittens, A., Achlioptas, D., and Mahoney, M. W. Skip-Gram - Zipf + Uniform = Vector Additivity. In *Association for Computational Linguistics*, 2017.

Hashimoto, T. B., Alvarez-Melis, D., and Jaakkola, T. S. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 2016.

Le, M. Unshifting the PMI matrix. `https://minhlab.wordpress.com/2017/02/16/presentation-at-clin-27/`, presented at CLIN 27 (2017), 2017. [Online; accessed Sep 2018, presented at CLIN 27, 2017].

Levy, O. and Goldberg, Y. Linguistic regularities in sparse and explicit word representations. In *Computational Natural Language Learning*, 2014a.

Levy, O. and Goldberg, Y. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2014b.

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

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013a.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013b.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, 2014.

Socher, R., Bauer, J., Manning, C. D., et al. Parsing with compositional vector grammars. In *Association for Computational Linguistics*, 2013.

Srebro, N. and Jaakkola, T. Weighted low-rank approximations. In *International Conference on Machine Learning*, 2003.

Turian, J., Ratinov, L., and Bengio, Y. Word representations: a simple and general method for semi-supervised learning. In *Association for Computational Linguistics*, 2010.

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