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* and show it can be re-interpreted as *word transformation*, a mathematical description of *"w_x is to w_y"*. From these concepts we prove existence of the linear relationship between W2V-type embeddings that underlies the analogical phenomenon, and identify explicit error terms in the relationship.

1. Introduction

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*[1] (W2V) (Mikolov et al., 2013a) and *Glove* (Pennington et al., 2014) are amongst the best known and on which we focus.

Interestingly, neural embeddings are observed to exhibit seemingly linear behaviour (e.g. Levy & Goldberg (2014a)). In particular, for *analogies*, or word relationships of the form *"w_a is to w_{a^*} as w_b is to w_{b^*}"*, the corresponding embeddings often satisfy $\mathbf{v}_{a*} - \mathbf{v}_a + \mathbf{v}_b \approx \mathbf{v}_{b*}$ (where $\mathbf{v}_i$ denotes an embedding of word w_i), enabling analogical questions such as "*man is to king as woman is to ..?*" to be solved by vector algebra, i.e. by completing the parallelogram. Such higher order structure is surprising since word embeddings are trained using only pairwise word co-occurrence data extracted from large corpora.

We begin by showing that *paraphrasing* determines when a linear combination of embeddings equates to that of another word for embeddings that factorise *pointwise mutual information* (PMI). By the paraphrase $king \approx_{\text{P}} \{man, royal\}$, we mean that $king$ is semantically equivalent to $\{man, royal\}$, where equivalence is measured in terms of the probability distributions induced over nearby words, fitting Firth's maxim "*You shall know a word by the company it keeps*" (Firth, 1957). We then show that paraphrases can be reinterpreted as *word transformations* 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^*} that take the same parameters, the linear relationship between word embeddings observed for analogies follows (See outline of steps in Fig 3). Our key contributions are:

- to derive a probabilistic definition of *paraphrasing* and prove that a word embedding relates to a sum of other embeddings according to their paraphrase relationship;
- to generalise paraphrasing to word sets and show such a paraphrase can be viewed as a *word transformation*, a mathematical description of *"w_x is to w_{x^*}"*;
- to provide the first rigorous explanation for the presence of linear relationships between word embeddings of analogies;
- to prove these relationships for embeddings that factorise PMI or of full-rank, or those from W2V or *Glove*; and
- to identify explicit, interpretable error terms in the linear relationships that relate to the strength of relevant paraphrases and statistical dependencies.

2. Previous Work

Intuition for the presence of linear analogical relationships, or *linguistic regularity*, in embeddings is provided by the authors of both W2V (Mikolov et al., 2013a;b) and *Glove*

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

Preprint. Work in progress.

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

(Pennington et al., 2014) and has been widely discussed (e.g. Levy & Goldberg (2014a); Linzen (2016)). More recently, several theoretical explanations have been proposed:

- The first rigorous attempt (Arora et al., 2016) proposes a latent variable model for language. This work contains several strong *a priori* assumptions regarding the spatial distribution of word vectors, as discussed by Gittens et al. (2017), that we do not require. Also, the two embedding weight matrices of W2V are assumed to be equal, which we show to be false in practice.

- An explanation by Gittens et al. (2017) uses *paraphrasing* from which we draw inspiration. However, several assumptions are relied on that do not hold in practice, notably: (i) that words follow a uniform distribution rather than their (non-uniform) Zipf distribution; (ii) that W2V learns a conditional distribution, which is violated due to negative sampling (Levy & Goldberg, 2014b); and (iii) that joint probabilities beyond pairwise co-occurrences are zero, which is empirically false.

- Recently, Ethayarajh et al. (2018) propose an explanation based on *co-occurrence shifted PMI*. However, that property itself is not motivated from word relationships; and several implicit assumptions, e.g. $\text{PMI}(w_i, w_i) = \log p(w_i)$ for word w_i, do not hold in practice.

Importantly, no previous work to our knowledge offers a mathematical interpretation of analogies and so an end-to-end explanation of why when *"w_a is to w_{a^*} as w_b is to w_{b^*}"* word embeddings exhibit a specific linear relationship.

3. Background

3.1. W2V

The W2V algorithm considers a set of word pairs $\{(w_{i_k}, w'_{j_k})\}_k$ generated from a (typically large) text corpus, by allowing the *target* word w_i to range over the corpus, and the *context* word w'_j to range over a symmetric context window (of size l) defined relative to 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 of W2V can be regarded more simply as matrix multiplication of 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 set of unique words in the corpus. By convention, $\mathbf{W}$ denotes the matrix closest to the input target words. Columns of $\mathbf{W}$ and $\mathbf{C}$ are the *embeddings* of words in $\mathcal{E}$: $\mathbf{v}_i$ (the i^{th} column of $\mathbf{W}$) corresponds to the i^{th} word in $\mathcal{E}$ as a target word, denoted w_i; and $\mathbf{v}'_i$ (the i^{th} column of $\mathbf{C}$) corresponds to the same word when observed as a context word, denoted w'_i.

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

$$\mathbf{v}_i^\top \mathbf{v}'_j \;=\; \text{PMI}(w_i, w'_j) - \log k\;, \tag{1}$$

where PMI refers to *pointwise mutual information*, defined as $\text{PMI}(w_i, w'_j) = \log \frac{p(w_i, w'_j)}{p(w_i)p(w'_j)}$. In matrix form, this corresponds to the implicit matrix factorisation:

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

where $\mathbf{SPMI}_{i,j} = \text{PMI}(w_i, w'_j) - \log k$, i.e. *shifted* PMI.

3.2. Glove

Embeddings learned by *Glove* (Pennington et al., 2014) perform comparably to those of W2V and exhibit similar linear structure for analogies. *Glove*'s architecture is the same as that of W2V but its loss function is optimised when:

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

where b_i, b_j are biases and Z a normalising constant. Eqn 3 generalises Eqn 1 due to the biases. As such, *Glove* has greater flexibility and a wider range of potential solutions. However, as we will show, it is the optimum of W2V's objective function (Eqn 1) that gives rise to the linear analogical structure of its embeddings and we conjecture that the same rationale underpins the linear analogical structure in *Glove* embeddings.

4. Preliminaries

The relationship between word embeddings and co-occurrence statistics (Eqns 1 and 2) is fundamental to the linear structure observed for analogies. We therefore highlight relevant aspects of the relationship.

4.1. Impact of the *Shift*

The direct appearance of k in the optimum of W2V's objective function (Eqn 1) and any associated impact on its embeddings is somewhat arbitrary since k reflects nothing of word properties. Also, the so-called *shift* $(-\log k)$ is material relative to empirical PMI values for typical values of k (Figure 1) and cannot simply be ignored. Further, it is known that adjusting the W2V algorithm to remove the direct impact of the *shift* improves embedding performance (Le, 2017). We conclude that the presence of the *shift* in Eqn 1 is a detrimental artefact of the W2V algorithm and so (unless stated otherwise) consider embeddings that factorise $\mathbf{PMI} \in \mathbb{R}^{n\times n}$, the *unshifted* PMI matrix, i.e.:

$$\begin{aligned} \mathbf{v}_i^\top \mathbf{v}'_j \;&=\; \text{PMI}(w_i, w'_j) \\ \text{or} \qquad \mathbf{W}^\top \mathbf{C} \;&=\; \mathbf{PMI}\;. \end{aligned} \tag{4}$$

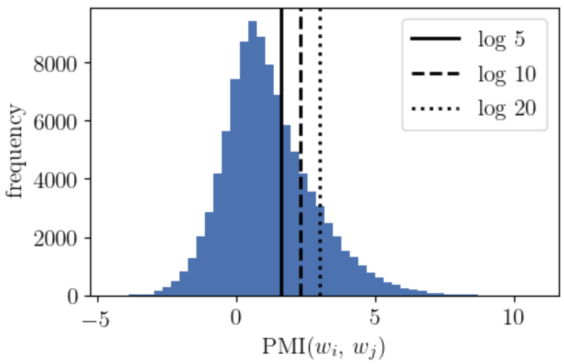


Figure 1. Histogram of PMI(w_i, w'_j) for word pairs w_i, w'_j extracted from sample text, with $\log k$ shown for typical k.

4.2. Reconstruction Error

In practice, Eqns 2 and 4 hold only approximately since $\mathbf{W}^\top\mathbf{C}$ (of dimension $n\times n$ with rank at most $d<n$) is a rank-constrained approximation. Thus recovering PMI values from $\mathbf{W}$ and $\mathbf{C}$ is subject to *reconstruction error*. However, we rely throughout on linear relationships in $\mathbb{R}^n$ and only require that these are maintained when projected (down) into the d-dimensional space of embeddings. To ensure this, we assume:

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

A2. $\mathbf{W}$ *is an accurate projection of the full rank matrix (e.g.* $\mathbf{PMI}$ *in Eqn 4) onto the rows of* $\mathbf{C}$*, i.e. by* $\mathbf{C}^\dagger = (\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C}$*, the Moore-Penrose pseudo inverse of* $\mathbf{C}^\top$.

A1 is reasonable since $d\ll n$ and d is chosen. A2 assumes that, whatever the factorisation method (e.g. analytic, W2V, *Glove* or otherwise, e.g. Srebro & Jaakkola (2003)), an accurate low-rank approximation is achieved.[2] Together, A1 and A2 imply that under projection by $\mathbf{C}^\dagger$, linear relationships between column vectors of $\mathbf{PMI}$ also hold for columns of $\mathbf{W}$, i.e. embeddings $\mathbf{v}_i$. The converse cannot be assumed.

4.3. Zero Counts of Word Co-occurrences

Co-occurrences of rare words may be unobserved, thus their empirical probability estimates are zero and corresponding PMI estimates undefined. However, such zero counts become less common for a fixed dictionary $\mathcal{E}$ as the corpus size or context window size l is increased, where the latter can be made arbitrarily large with down-weighting applied to more distant words (e.g. Pennington et al. (2014)). In this work, we consider small word sets $\mathcal{W}$ and assume the corpus and context window to be of sufficient relative size that *true* probabilities considered are non-zero and corresponding PMI values well-defined, i.e.:

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

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

4.4. 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. $\mathbf{v}_i=\mathbf{v}'_i\ \forall i$. The assumption is convenient as the number of parameters is halved, equations simplify and consideration of how to combine $\mathbf{v}_i$ and $\mathbf{v}'_i$ for downstream use falls away. However, the assumption implies $\mathbf{W}^\top\mathbf{W}=\mathbf{PMI}$ that requires $\mathbf{PMI}$ to be positive semi-definite, which is not true for typical corpora. Thus $\mathbf{v}_i$, $\mathbf{v}'_i$ are not equal and modifying W2V to enforce them to be equal 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 $\mathbf{v}_\mathcal{W}=\sum_i\mathbf{v}_i$. 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 together, w_* could appear instead. This suggests a relationship between the probability distributions $p(w'_j|\mathcal{W})$ and $p(w'_j|w_*)$, $\forall w'_j\in\mathcal{E}$. We refer to such a conditional distribution over all context words as the distribution *induced* by $\mathcal{W}$ or w_*, respectively.

5.1. Defining a Paraphrase

Letting $\mathcal{C}_\mathcal{W}=\{w'_{j_1},\dots,w'_{j_t}\}$ denote the 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 $w_*^{(1)}=\text{argmax}_{w_i\in\mathcal{E}}\,p(\mathcal{C}_\mathcal{W}|w_i)$. Assuming $w'_j\in\mathcal{C}_\mathcal{W}$ to be independent draws from $p(w'_j|\mathcal{W})$, gives:

$$\begin{aligned} w_*^{(1)} &= \operatorname*{argmax}_{w_i} \textstyle\prod_{w'_j\in\mathcal{E}} p(w'_j|w_i)^{\#_j} \\ &\longrightarrow \operatorname*{argmax}_{w_i} \textstyle\sum_{w'_j\in\mathcal{E}} p(w'_j|\mathcal{W})\log p(w'_j|w_i)\,, \end{aligned}$$

as $|\mathcal{C}_\mathcal{W}|\longrightarrow\infty$, where $\#_j$ denotes the count of w'_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.:

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

Alternatively, we might consider $w_*^{(2)}$, the target word whose set of associated context words $\mathcal{C}_{w_*}$ is best explained by $\mathcal{W}$

[2] We note that the objective function of W2V has other possible solutions, which may obfuscate the relationships we identify.

[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.

in the sense that $w_*^{(2)}$ minimises KL divergence $\Delta_{KL}^{w_*,\mathcal{W}} = D_{KL}[P(w'_j|w_*)\,||\,P(w'_j|\mathcal{W})]$ (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 achieves its lower bound (zero) *iff* the induced distributions are the same, 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:*

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

Note that $\mathcal{W}$ and w_* need not appear similarly often for w_* to paraphrase $\mathcal{W}$, only amongst the same context words.

5.2. Paraphrase = Embedding Sum + Error

We now show how paraphrase error arises within word embedding relationships, specifically between $\mathbf{v}_*$ and $\mathbf{v}_\mathcal{W}$.

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 $\boldsymbol{\sigma}_j^{\mathcal{W}} = \log \frac{p(\mathcal{W}|w'_j)}{\prod_i p(w_i|w'_j)}$ and $\tau^{\mathcal{W}} = \log \frac{p(\mathcal{W})}{\prod_i p(w_i)}$.

Proof. (See Appendix B.) □

Since Lem 1 underpins all that follows, we provide a brief commentary on its proof. A correspondence is drawn between (I) the product of distributions induced by each $w_i \in \mathcal{W}$ and (II) the distribution induced by w_* (which may be an entirely unrelated word) by comparing each to (III) the distribution induced by joint event $\mathcal{W}$, i.e. observing all $w_i \in \mathcal{W}$ simultaneously. I closely compares to III if $w_i \in \mathcal{W}$ are independent, as reflected by $\boldsymbol{\sigma}_j^{\mathcal{W}}$, $\tau^{\mathcal{W}}$.[4] II relates to III by the paraphrase error $\boldsymbol{\rho}_j^{\mathcal{W},w_*}$. From Lem 1 we obtain:

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

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

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

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

[4] Analogously to a product of marginal distributions equating to the corresponding joint probability subject to independence.

Thm 1 shows that the difference between any embedding $\mathbf{v}_*$ and a sum of embeddings $\mathbf{v}_\mathcal{W}$ can be attributed to the paraphrase error $\boldsymbol{\rho}^{\mathcal{W},w_*}$ between w_* and $\mathcal{W}$, and terms $\boldsymbol{\sigma}^{\mathcal{W}}$ and $\tau^{\mathcal{W}}$ that reflect relationships within $\mathcal{W}$ (independent of w_*), referred to collectively as *dependence error*:

- $\boldsymbol{\sigma}^{\mathcal{W}}$ is a vector reflecting conditional dependencies within $\mathcal{W}$ given each $w'_j \in \mathcal{E}$; $\boldsymbol{\sigma}_j^{\mathcal{W}} = 0$ *iff* conditional independence holds over all $w_i \in \mathcal{W}$ given $w'_j \in \mathcal{E}$;
- $\tau^{\mathcal{W}}$ is a scalar measure of mutual independence of $w_i \in \mathcal{W}$ (thus constant $\forall w'_j \in \mathcal{E}$); $\tau^{\mathcal{W}} = 0$ *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 $w'_j \in \mathcal{E}$.*

Thm 1, which holds for all words w_* and word sets $\mathcal{W}$, explains why and when a paraphrase (e.g. $\{man, royal\}$ by $king$) can be identified by embedding addition ($\mathbf{v}_{man} + \mathbf{v}_{royal} \approx \mathbf{v}_{king}$). 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 $\mathbf{v}_* - \mathbf{v}_\mathcal{W}$ 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 $\mathcal{W} \subseteq \mathcal{E}$, $\mathbf{v}_* \approx \mathbf{v}_\mathcal{W}$ 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: $\mathbf{v}_* \approx \mathbf{v}_\mathcal{W}$ does not imply w_* paraphrases $\mathcal{W}$. It can be seen that *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 so project to an overall small vector difference between $\mathbf{v}_*$ and $\mathbf{v}_\mathcal{W}$.

In principle, case (i) can arise in both PMI vectors (Lem 1) and lower rank embeddings (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 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 $\mathbf{W} = \mathbf{PMI}$, $\mathbf{C} = \mathbf{I}$ (identity matrix) in Thm 1. □

5.5. Paraphrasing in W2V Embeddings

Thm 1 can be extended to W2V embeddings by substituting $\mathbf{v}_i{}^\top \mathbf{v}'_j = \text{PMI}(w_i, w'_j) - \log k$ appropriately to obtain:

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

$$\mathbf{v}_* = \mathbf{v}_\mathcal{W} + \mathbf{C}^\dagger(\boldsymbol{\rho}^{\mathcal{W},w_*} + \boldsymbol{\sigma}^\mathcal{W} - \tau^\mathcal{W}\mathbf{1} + \log k(|\mathcal{W}|-1)\mathbf{1})\,. \tag{7}$$

Comparing Eqns 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 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:

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

or why in the more general case:

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

$\forall (w_x, w_{x^*}) \in S_\mathfrak{A}$ 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 Eqns 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.

6.1. Paraphrasing Word Sets

We first generalise the notion of paraphrasing to word sets:

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

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

D2 generalises D1 such that the paraphrase term $\mathcal{W}_*$, previously w_*, can be more than one word.[5] Analogously to before (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.) □

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

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

Proof. Multiply Eqn 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 $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$, $|\mathcal{W}^+|, |\mathcal{W}^-| < l-1$:*

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

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 $\mathbf{v}_{x^*}$ relates to a linear combination of other embeddings ($\mathbf{v}_\Sigma = \mathbf{v}_x + \mathbf{v}_{\mathcal{W}^+} - \mathbf{v}_{\mathcal{W}^-}$), following an equivalent relationship between columns of $\mathbf{PMI}$. Analogously to single-word (or *direct*) paraphrases, the vector difference $\mathbf{v}_{x^*} - \mathbf{v}_\Sigma$ depends on the paraphrase error that reflects the relationship between the two word sets $\mathcal{W}_*$ and $\mathcal{W}$; and the dependence error that reflects statistical dependence of words within each of $\mathcal{W}$ and $\mathcal{W}_*$.

Corollary 2.2. *For terms as defined above, $\mathbf{v}_{x^*} \approx \mathbf{v}_x + \mathbf{v}_{\mathcal{W}^+} - \mathbf{v}_{\mathcal{W}^-}$ 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.

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

6.2. From Paraphrases to Analogies

A special case of Cor 2.1 gives:

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

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

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. □

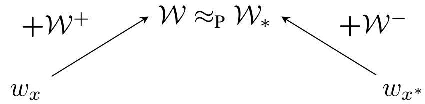
Thus we see that Eqn 8 holds *if* $\{w_{b^*}, w_a\} \approx_{\text{P}} \{w_b, w_{a^*}\}$ and those word sets exhibit *similar dependence* (Sec 6.6). More generally, we see from Cor 2.1 that Eqn 9 is satisfied with $\mathbf{u}_{\mathfrak{A}} \approx \mathbf{v}_{\mathcal{W}^+} - \mathbf{v}_{\mathcal{W}^-}$ *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 sets exhibit similar dependence.

This establishes sufficient conditions for the word embedding relationships observed for analogies (Eqns 8 and 9) based on interpretable word relationships, answering Q1. However, these relationships are paraphrases that show no immediate connection to the "w_x *is to* w_{x^*}..." relationships of analogies we aim for. To prove that the requisite paraphrases follow directly from analogical relationships we introduce the concept of *word transformation*.

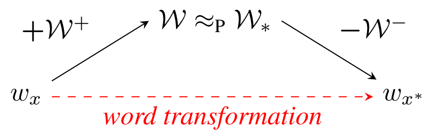
6.3. Word Transformation

Thus far, the paraphrase of a word set $\mathcal{W}$ by word w_* (D1) has been considered in the sense of an equivalence between $\mathcal{W}$ and w_* in terms of their induced distributions. Alternatively, such a paraphrase can be interpreted as a *transformation* from a particular word $w_s \in \mathcal{W}$ 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 a distribution 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 more closely resembling that of w_*. This mirrors a similar view of the associated embedding addition: starting with $\mathbf{v}_s$ add $\mathbf{v}_i$, $\forall w_i \in \mathcal{W}^+$, to approximate $\mathbf{v}_*$. Note that only *addition* applies.

Extending this to D2, The paraphrasing of word set $\mathcal{W}$ by another $\mathcal{W}_*$ can be interpreted additively as starting with $w_x \in \mathcal{W}$ and $w_{x^*} \in \mathcal{W}_*$, 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 one another. Context is added to both w_x and w_{x^*} until contextualised cases $\mathcal{W}$ and $\mathcal{W}_*$ paraphrase (Fig 2a). Note $\mathcal{W}$ and $\mathcal{W}_*$ may have no intuitive meaning and need not correspond to a single word, unlike D1 paraphrases.



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

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^*}, within which the paraphrase between $\mathcal{W}$ and $\mathcal{W}_*$ effectively serves as an intermediate step (Fig 2b). 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 with D1 paraphrases, i.e. by including *differences* between words. More precisely, transformation parameters align the induced distributions and create a paraphrase.

This interpretation shows an 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 removal of context that mirrors 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.

Definition D 3. *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 the 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**"

Having word transformation as a means to describe the semantic difference between words, we mathematically inter-

[6] Analogously to standard algebra: if $x > y$, equality is achieved by subtracting from x: $x - d = y$; or by adding to y: $a = y + d$ (where x, y and d are all positive values).

pret 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 specify 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 anticipated linear relationships between word embeddings (Eqns 8 and 9) follow from D4.

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

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

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. Substitute $\mathcal{W} = \mathcal{W}^x$, $\mathcal{W}_* = \mathcal{W}^x_*$ for $x \in \{a, b\}$ into 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. □

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

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

where terms are as defined in Lem 3.

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

More generally, if D4 holds $\forall\, (w_a, w_{a^*}), (w_b, w_{b^*}) \in S$, a set of ordered word pairs, and $\mathcal{W}^+$, $\mathcal{W}^-$ simultaneously transform w_x to w_{x^*}, $\forall (w_x, w_{x^*}) \in S$, then each set $\{w_{x^*}, \mathcal{W}^-\}$ must paraphrase $\{w_x, \mathcal{W}^+\}$ (by 3) and Eqn 11 holds with small paraphrase error. By this and Thm 3 we know that the word embeddings of an analogy satisfy linear relationships (Eqns 8 and 9), subject to dependence error.

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

6.4.1. 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:

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

Eqn 13 also applies (by suitably redefining $\mathcal{W}^x$, $\mathcal{W}^x_*$) if there exist other parameters $\mathcal{W}^+_2, \mathcal{W}^-_2$ that (*w.l.o.g.*) transform w_a to w_{a^*}; whereby $\boldsymbol{\rho}^{\mathcal{W}^a, \mathcal{W}^a_*}$ is small but nothing is known of $\boldsymbol{\rho}^{\mathcal{W}^b, \mathcal{W}^b_*}$. Thus, subject to dependence error:

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

By Eqns 14 and 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 any such 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^*} $\forall (w_x, w_{x^*}) \in S_{\mathfrak{A}}$.

6.4.2. 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 such choice simplifies Thm 3 to:

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

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

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

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

We arrive back at Eqn 12 but now link directly to analogies, proving that word embeddings of analogies satisfy linear relationships (Eqns 8 and 9), subject to dependence error. Fig 3 shows a summary of all steps to prove Cor 3.2. By D4, we also have a mathematical description 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, $\mathfrak{A}^*$ implies 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 (arbitrarily) as $\mathcal{W}^+ = \{queen\}, \mathcal{W}^- = \{woman\}$. 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:

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

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

$$
\begin{array}{ccccccc}
\text{“}w_a \textit{ is to } w_{a^*} & & w_a \underset{\mathcal{W}^-}{\overset{\mathcal{W}^+}{\longrightarrow}} w_{a^*} & & \{w_a, \mathcal{W}^+\} \approx_{\mathrm{P}} \{w_{a^*}, \mathcal{W}^-\} & & \mathbf{v}_{a^*} - \mathbf{v}_a \\
\textit{as} & \iff & \wedge & \iff & \wedge & \implies & \approx \\
w_b \textit{ is to } w_{b^*}\text{”} & & w_b \underset{\mathcal{W}^-}{\overset{\mathcal{W}^+}{\longrightarrow}} w_{b^*} & & \{w_b, \mathcal{W}^+\} \approx_{\mathrm{P}} \{w_{b^*}, \mathcal{W}^-\} & & \mathbf{v}_{b^*} - \mathbf{v}_b
\end{array}
$$

Figure 3. Summary of steps to prove the relationship between analogies and word embeddings (omitting dependence error). $w_x \underset{\mathcal{W}^-}{\overset{\mathcal{W}^+}{\longrightarrow}} w_{x^*}$ denotes a word transformation w_x to w_{x^*} with parameters $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$.

where words are abbreviated by their initials and, explicitly:

$$
\begin{aligned}
\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} &= \log \tfrac{p(w'_j|w_Q,w_M)}{p(w'_j|w_W,w_K)} && \text{(which must be small)},\\
\boldsymbol{\sigma}^{\mathcal{W}} &= \log \tfrac{p(w_W,w_K|w'_j)}{p(w_W|w'_j)p(w_K|w'_j)}, & \tau^{\mathcal{W}} &= \log \tfrac{p(w_W,w_K)}{p(w_W)p(w_K)},\\
\boldsymbol{\sigma}^{\mathcal{W}_*} &= \log \tfrac{p(w_Q,w_M|w'_j)}{p(w_Q|w'_j)p(w_M|w'_j)}, & \tau^{\mathcal{W}_*} &= \log \tfrac{p(w_Q,w_M)}{p(w_Q)p(w_M)}\ .
\end{aligned}
$$

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

6.6. Dependence error in analogies

Dependence error terms for analogies (e.g. Eqn 16) bear an important distinction from those in general paraphrases (Eqn 6). 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 and can only cancel one another by chance (which will be low in high dimensions). However, Eqn 16 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, 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 analogy "w_a is to w_{a^*} as w_a is to w_{a^*}", which is trivially true with zero total error. Here, paraphrase error is zero since $p(w'_j|\{w_a, w_{a^*}\}) = p(w'_j|\{w_{a^*}, w_a\})$, $\forall w'_j \in \mathcal{E}$, and 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 trivial example 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* $\mathbf{PMI}$ *by setting* $\mathbf{C}=\mathbf{I}$.

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, i.e.:

Corollary 3.4. *Cor 3.2 applies to W2V embeddings.*

Thus, linear relationships between embeddings for analogies hold equally for W2V embeddings as for those derived without the *shift* distortion. This is perhaps surprising but is corroborative since linear analogical relationships have been observed extensively in W2V embeddings (e.g. Levy & Goldberg (2014a)), as is now theoretically justified. Thus we know that analogies hold for W2V embeddings subject to higher order statistical relationships between words of the analogy 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 show such paraphrase relationships to be fundamental to linear relationships between word embeddings that factorise PMI or similar (e.g. W2V, Glove or full rank columns of the PMI matrix). We further show that paraphrases can be reinterpreted as word transformations, enabling analogies to be mathematically defined. Thus, by translating differences in semantics to differences between embeddings we provide the first rigorous explanation for linear relationships between the word embeddings of analogies.

In future work we will look to extend our understanding of the relationships between word embeddings to other applications of discrete object representation, e.g. graph embedding and recommender systems, that rely on an underlying matrix factorisation. Furthermore, 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.

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 *Proceedings of the 55th Annual Meeting of the 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 *Proceedings of the 18th conference on 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. *arXiv preprint arXiv:1606.07736*, 2016.

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 *Proceedings of the 2013 Conference of the 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 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Socher, R., Bauer, J., Manning, C. D., et al. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, 2013.

Srebro, N. and Jaakkola, T. Weighted low-rank approximations. In *Proceedings of the 20th 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 *Proceedings of the 48th Annual Meeting of the 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.

Appendices

A. The KL-divergence between induced distributions

We consider the words found by minimising the difference KL-divergences considered in Section 5. Specifically:

$$w_*^{(1)} = \operatorname*{argmin}_{w_i \in \mathcal{E}} D_{KL}[\,p(w'_j|\mathcal{W})\,||\,p(w'_j|w_i)\,]$$
$$w_*^{(2)} = \operatorname*{argmin}_{w_i \in \mathcal{E}} D_{KL}[\,p(w'_j|w_i)\,||\,p(w'_j|\mathcal{W})\,]$$

Minimising $D_{KL}[\,p(w'_j|\mathcal{W})\,||\,p(w'_j|w_i)\,]$ identifies the word that induces a probability distribution over context words closest to that induced by $\mathcal{W}$, in which probability mass is assigned to w'_j *wherever* it is for $\mathcal{W}$. Intuitively, $w_*^{(1)}$ is the word that most closely reflects *all* aspects of $\mathcal{W}$, and may occur in contexts where no word $w_i \in \mathcal{W}$ does.

Minimising $D_{KL}[\,p(w'_j|w_i)\,||\,p(w'_j|\mathcal{W})\,]$ finds the word that induces a distribution over context words that is closest to that induced by $\mathcal{W}$, in which probability mass is assigned as broadly as possible but *only* to those w'_j to which probability mass is assigned for $\mathcal{W}$. Intuitively, $w_*^{(2)}$ is the word that reflects as many aspects of $\mathcal{W}$ as possible, as closely as possible, but nothing additional, e.g. by having other meaning that $\mathcal{W}$ does not.

A.1. Weakening the paraphrase assumption

For a given word set $\mathcal{W}$, we consider the relationship between embedding sum $\mathbf{v}_\mathcal{W}$ and embedding $\mathbf{v}_*$ for the word $w_* \in \mathcal{E}$ that minimises the KL-divergence (we illustrate with $\Delta_{KL}^{\mathcal{W},w_*}$). Exploring a weaker assumption than D1, tests whether D1 might exceed requirement, and explores the relationship between $\mathbf{v}_*$ and $\mathbf{v}_\mathcal{W}$ as paraphrase error increases.

Theorem 4 (Weak paraphrasing). *For $w_* \in \mathcal{E}, \mathcal{W} \subseteq \mathcal{E}$,* if *$w_*$ minimises $\Delta_{KL}^{\mathcal{W},w_*} \doteq D_{KL}[\,p(w'_j|\mathcal{W})\,||\,p(w'_j|w_*)\,]$, then:*

$$\mathbf{v}_*{}^\top \hat{\mathbf{v}}' \;=\; \mathbf{v}_\mathcal{W}{}^\top \hat{\mathbf{v}}' - \Delta_{KL}^{\mathcal{W},w_*} + \hat{\sigma}^\mathcal{W} - \tau^\mathcal{W} \qquad (17)$$

where $\hat{\mathbf{v}}' = \mathbb{E}_{j|\mathcal{W}}[\mathbf{v}'_j]$, $\hat{\sigma}^\mathcal{W} = \mathbb{E}_{j|\mathcal{W}}[\boldsymbol{\sigma}_j^\mathcal{W}]$ and $\mathbb{E}_{j|\mathcal{W}}[\cdot]$ denotes expectation under $p(w'_j|\mathcal{W})$.

Proof.

$$\begin{aligned}\Delta_{KL}^{\mathcal{W},w_*} &= \textstyle\sum_j p(w'_j|\mathcal{W}) \log \frac{p(w'_j|\mathcal{W})}{p(w'_j|w_*)} \\ &\overset{(5)}{=} \mathbb{E}_{j|\mathcal{W}}[\,\textstyle\sum_i \mathrm{PMI}(w_i, w'_j) \\ &\qquad - \mathrm{PMI}(w_*, w'_j) + \boldsymbol{\sigma}_j^\mathcal{W} - \tau^\mathcal{W}] \\ &= \mathbb{E}_{j|\mathcal{W}}[\mathbf{v}_\mathcal{W}{}^\top \mathbf{v}'_j - \mathbf{v}_*{}^\top \mathbf{v}'_j] + \hat{\sigma}^\mathcal{W} - \tau^\mathcal{W} \qquad \square\end{aligned}$$

Thus, the weaker paraphrase relationship specifies a hyperplane containing $\mathbf{v}_*$ and so does not uniquely define $\mathbf{v}_*$ (as under D1) and cannot explain the observation of embedding addition for paraphrases (as suggested by Gittens et al. (2017)). A similar result holds for $\Delta_{KL}^{w_*,\mathcal{W}}$. In principle, Thm 4 could help locate embeddings of words that more loosely paraphrase $\mathcal{W}$, i.e. with increased paraphrase error.

B. Proof of Lemma 1

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 $\boldsymbol{\sigma}_j^\mathcal{W} = \log \frac{p(\mathcal{W}|w'_j)}{\prod_i p(w_i|w'_j)}$ and $\tau^\mathcal{W} = \log \frac{p(\mathcal{W})}{\prod_i p(w_i)}$.

Proof.

$$\begin{aligned}&\mathrm{PMI}(w_*, w'_j) - \sum_{w_i \in \mathcal{W}} \mathrm{PMI}(w_i, w'_j) \\ &= \log \frac{p(w_*|w'_j)}{p(w_*)} - \log \prod_{w_i \in \mathcal{W}} \frac{p(w_i|w'_j)}{p(w_i)} \\ &= \log \frac{p(w_*|w'_j)}{\prod_\mathcal{W} p(w_i|w'_j)} - \log \frac{p(w_*)}{\prod_\mathcal{W} p(w_i)} \\ &\qquad + \log \frac{p(\mathcal{W}|w'_j)}{p(\mathcal{W}|w'_j)} + \log \frac{p(\mathcal{W})}{p(\mathcal{W})} \\ &= \log \frac{p(w_*|w'_j)}{p(\mathcal{W}|w'_j)} - \log \frac{p(w_*)}{p(\mathcal{W})} \\ &\qquad + \log \frac{p(\mathcal{W}|w'_j)}{\prod_\mathcal{W} p(w_i|w'_j)} - \log \frac{p(\mathcal{W})}{\prod_\mathcal{W} p(w_i)} \\ &= \log \frac{p(w'_j|w_*)}{p(w'_j|\mathcal{W})} + \log \frac{p(\mathcal{W}|w'_j)}{\prod_\mathcal{W} p(w_i|w'_j)} \\ &\qquad - \log \frac{p(\mathcal{W})}{\prod_\mathcal{W} p(w_i)} \\ &= \boldsymbol{\rho}_j^{\mathcal{W},w_*} + \boldsymbol{\sigma}_j^\mathcal{W} - \tau^\mathcal{W}\,,\end{aligned}$$

where, unless stated explicitly, products are with respect to all w_i in the set indicated. $\square$

Introduced terms are highlighted to show their evolution within the proof. At the step where terms are introduced, the existing error terms have no statistical meaning. This is resolved by introducing terms to which both error terms can be meaningfully related, through paraphrasing and independence.

C. Proof of Lemma 2

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

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

Proof.

$$\begin{aligned}
&\sum_{w_i \in \mathcal{W}_*} \text{PMI}(w_i, w'_j) - \sum_{w_i \in \mathcal{W}} \text{PMI}(w_i, w'_j) \\
&= \log \prod_{w_i \in \mathcal{W}_*} \frac{p(w_i|w'_j)}{p(w_i)} - \log \prod_{w_i \in \mathcal{W}} \frac{p(w_i|w'_j)}{p(w_i)} \\
&= \log \frac{\prod_{\mathcal{W}_*} p(w_i|w'_j)}{\prod_{\mathcal{W}}\, p(w_i|w'_j)} - \log \frac{\prod_{\mathcal{W}_*} p(w_i)}{\prod_{\mathcal{W}}\, p(w_i)} \\
&\quad + \log \frac{p(\mathcal{W}_*|w'_j)}{p(\mathcal{W}_*|w'_j)} + \log \frac{p(\mathcal{W}_*)}{p(\mathcal{W}_*)} \\
&\quad + \log \frac{p(\mathcal{W}|w'_j)}{p(\mathcal{W}|w'_j)} + \log \frac{p(\mathcal{W})}{p(\mathcal{W})} \\
&= + \log \frac{p(\mathcal{W}_*|w'_j)}{p(\mathcal{W}|w'_j)} - \log \frac{p(\mathcal{W}_*)}{p(\mathcal{W})} \\
&\quad + \log \frac{\prod_{\mathcal{W}_*} p(w_i|w'_j)}{p(\mathcal{W}_*|w'_j)} - \log \frac{\prod_{\mathcal{W}_*} p(w_i)}{p(\mathcal{W}_*)} \\
&\quad + \log \frac{p(\mathcal{W}|w'_j)}{\prod_{\mathcal{W}} p(w_i|w'_j)} - \log \frac{p(\mathcal{W})}{\prod_{\mathcal{W}} p(w_i)} \\
&= + \log \frac{p(w'_j|\mathcal{W}_*)}{p(w'_j|\mathcal{W})} \\
&\quad + \log \frac{p(\mathcal{W}|w'_j)}{\prod_{\mathcal{W}} p(w_i|w'_j)} - \log \frac{p(\mathcal{W}_*|w'_j)}{\prod_{\mathcal{W}_*} p(w_i|w'_j)} \\
&\quad - \log \frac{p(\mathcal{W})}{\prod_{\mathcal{W}} p(w_i)} + \log \frac{p(\mathcal{W}_*)}{\prod_{\mathcal{W}_*} p(w_i)} \\
&= \boldsymbol{\rho}_j^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}_j^{\mathcal{W}} - \boldsymbol{\sigma}_j^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\,,
\end{aligned}$$

where, unless stated explicitly, products are with respect to all w_i in the set indicated. □

The proof is analogous to that of Lem 1, with more terms added (as highlighted) to an equivalent effect. A key difference to single-word (or *direct*) paraphrases (D1) is that the paraphrase is between two word sets $\mathcal{W}$ and $\mathcal{W}_*$ that need not correspond to any single word. The paraphrase error $\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*}$ compares the induced distributions of the two sets, following the same principles as direct paraphrasing, but with perhaps less interpretatability.

D. Alternate Proof of Corollary 2.1

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

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

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

Proof.

$$\begin{aligned}
&\text{PMI}(w_{x^*}, w'_j) - \text{PMI}(w_x, w'_j) \\
&= \log \frac{p(w'_j|w_{x^*})}{p(w'_j|w_x)} + \log \prod_{w_i \in \mathcal{W}^+} \frac{p(w'_j|w_i)}{p(w'_j|w_i)} \\
&\qquad + \log \prod_{w_i \in \mathcal{W}^-} \frac{p(w'_j|w_i)}{p(w'_j|w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \log p(w'_j|w_i) - \sum_{w_i \in \mathcal{W}^-} \log p(w'_j|w_i) \\
&\qquad + \log \frac{\prod_{\mathcal{W}_*} p(w'_j|w_i)}{\prod_{\mathcal{W}}\, p(w'_j|w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \text{PMI}(w_i, w'_j) - \sum_{w_i \in \mathcal{W}^-} \text{PMI}(w_i, w'_j) \\
&\qquad + \log \frac{\prod_{\mathcal{W}_*} p(w_i|w'_j) \prod_{\mathcal{W}}\, p(w_i)}{\prod_{\mathcal{W}}\, p(w_i|w'_j) \prod_{\mathcal{W}_*} p(w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \text{PMI}(w_i, w'_j) - \sum_{w_i \in \mathcal{W}^-} \text{PMI}(w_i, w'_j) \\
&\qquad + \log \frac{p(w'_j|w_{x^*}, W^-)}{p(w'_j|w_x,\, W^+)} \\
&\qquad + \log \frac{\prod_{\mathcal{W}_*} p(w_i|w'_j)}{p(w_{x^*}, W^-|w'_j)} \frac{p(w_x, W^+|w'_j)}{\prod_{\mathcal{W}} p(w_i|w'_j)} \\
&\qquad - \log \frac{\prod_{\mathcal{W}_*} p(w_i)}{p(w_{x^*}, W^-)} \frac{p(w_x, W^+)}{\prod_{\mathcal{W}} p(w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \text{PMI}(w_i, w'_j) - \sum_{w_i \in \mathcal{W}^-} \text{PMI}(w_i, w'_j) \\
&\qquad + \boldsymbol{\rho}_j^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}_j^{\mathcal{W}} - \boldsymbol{\sigma}_j^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\,,
\end{aligned}$$

where, unless stated explicitly, products are with respect to all w_i in the set indicated; and $\mathcal{W} = \{w_x\} \cup \mathcal{W}^+$, $\mathcal{W}_* = \{w_{x^*}\} \cup \mathcal{W}^-$ to lighten notation. Multiplying by $\mathbf{C}^\dagger$ completes the proof. □