

On the Universality of the Inner Product as a PMI Estimator in Embedding Models

Enrique Amigó and Fernando Giner and Julio Gonzalo

ETSI Informática UNED, nlp.uned.es
Juan del Rosal, 16, 28040 Madrid, Spain
{enrique,fginer,julio}@lsi.uned.es

Abstract

Embedding representations are widely used in Natural Language Processing, yet their semantic interpretation depends on the operators used to compare them. Previous work has shown that, for certain embedding models, the inner product is linearly correlated with Pointwise Mutual Information (PMI). However, this relation has so far been established, empirically or analytically, only for certain specific models.

In this work, we formally show that, under certain assumptions about distributional semantic embedding models, the inner product between embeddings approximates PMI. This theoretical result provides a more general explanation for the observed relationship and enables the formulation of training criteria aimed at learning embedding spaces in which operations between embeddings are semantically interpretable.

1 Introduction

Embeddings currently play a central role in Natural Language Processing, but their effective training and use depend on the definition of appropriate embedding operators, such as similarity, entailment, and composition. Within Distributional Semantics, where expressions are considered semantically similar when they occur in similar contexts, estimating Pointwise Mutual Information (PMI) from embeddings is especially important, since PMI captures the extent to which two expressions share contexts beyond what would be expected by chance. This makes PMI estimation a cornerstone for defining such operators (see Section 5).

In this work, we formally show that, under certain assumptions about distributional semantic embedding models, the inner product between embeddings approximates PMI. That is, given the embeddings $\mathbf{v}_x$ and $\mathbf{v}_y$ associated with the textual expressions x and y :

$$\mathbf{v}_x \cdot \mathbf{v}_y \simeq \text{PMI}(x, y).$$

This correspondence has been analytically derived for Skip-gram with Negative Sampling (SGNS) embeddings by [Levy and Goldberg \(2014\)](#). In [Arora et al. \(2016\)](#) the analysis is extended to generative models under the assumption of isotropy.

There is also indirect empirical evidence. For instance, this correspondence follows that the self-information ($\text{PMI}(x, x) = -\log(P(x))$) is linearly correlated to the squared norm of the embedding: (i.e., $\mathbf{v}_x \cdot \mathbf{v}_x = \|\mathbf{v}_x\|^2$), and [Oyama et al. \(2023\)](#) demonstrated that the squared norm of word embeddings are effective indicators of word informativeness. Similarly, [Yokoi et al. \(2020\)](#) proposed a decomposition of word embeddings into norm and angle, where the norm reflects how much a word contributes to the overall meaning of a sentence. Other empirical studies have also reported a consistent correlation between embedding norms and information content in contextual language models ([Li et al., 2020](#); [Gao et al., 2019](#)).

All this empirical and analytical evidence, however, remains tied to specific embedding models. Our contribution, in contrast, is to formally prove that, under certain assumptions about distributional semantic embedding models, the inner product between embeddings approximates PMI. This theoretical result provides a more general explanation for the observed relationship, and enables the formulation of training criteria aimed at learning embedding spaces in which operations between embeddings are semantically interpretable.

2 Unified Probability Space for Embeddings and Word Sequences

Our goal is to study the relationship between the Pointwise Mutual Information (PMI) of linguistic expressions and the inner product of their corresponding embeddings. Since PMI is defined in terms of the marginal probabilities $P(x)$ and $P(y)$ and the joint probability $P(x, y)$, we first need to

establish a probability space in which these quantities are well defined.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, where the sample space Ω denotes the universe of possible *contexts*. We deliberately use the notion of context at an abstract level: an elementary context $\omega \in \Omega$ represents a possible contextual condition under which linguistic expressions may occur or be associated. In this way, contexts constitute the elementary outcomes of the probability space, while expressions will be represented as events over this space.

We do not commit to a single concrete notion of *context*, since its precise interpretation may depend on the training model under consideration. For instance, contexts may correspond to documents, as in Latent Semantic Indexing, to surrounding words or linguistic environments, as in SGNS, or to preceding token sequences, as in neural language models. At a more abstract level, context may also refer to semantic or pragmatic conditions under which an expression can occur or be interpreted. What is essential for our framework is only that contexts define the elementary outcomes over which the occurrence of expressions can be probabilistically characterized.

Let X be the set of expressions to be represented in the embedding space. We identify each expression $x \in X$ with an event in $\mathcal{F}$, namely, the set of contexts with which x is associated. Accordingly, $P(x)$ denotes the probability mass of the contexts associated with x . Given two expressions $x, y \in X$, their intersection corresponds to the contexts shared by both expressions, and therefore $P(x, y)$ denotes the probability mass of their shared contexts.

Notice that this notion of joint occurrence is defined over the abstract contextual space and does not necessarily imply that x and y cooccur within the same textual sequence. Depending on the particular notion of context adopted by a model, it may capture textual cooccurrence, shared linguistic environments, or more abstract semantic or pragmatic contextual conditions.

Moreover, we denote by $x_\top$ an expression carrying no information, that is, an expression associated with every possible context and satisfying $P(x_\top) = 1$. A typical example, under an appropriate definition of context, is a beginning-of-text token. We denote its corresponding embedding by $\mathbf{v}_\top$.

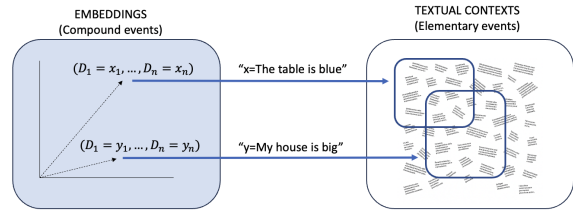


Figure 1: The probabilistic space. Contexts are elementary outcomes, while expressions and their embeddings are compound events defined over the contextual space. The joint probability $P(x, y)$ corresponds to the probability mass of the contexts shared by x and y .

We now connect this probabilistic framework with the embedding space. Let $\mathbf{v}_x = (x_1, \dots, x_n)$ denote the embedding associated with expression x . Assuming that the embedding map is injective over the set of expressions under consideration, we can induce a probability distribution over embeddings by defining $P(\mathbf{v}_x) = P(x)$. Thus, $P(\mathbf{v}_x)$ inherits the probabilistic interpretation of expression x , namely, the probability mass of the contexts associated with x ¹.

In addition, since an embedding corresponds to an n -dimensional vector, we can express $P(\mathbf{v}_x)$ as

$$P(x) = P(\mathbf{v}_x) = P(D_1 = x_1, \dots, D_n = x_n).$$

The fact that a dimension D_i takes a value x_i also constitutes a compound probabilistic event defined by the set of contexts associated with expressions whose value for dimension D_i is x_i . In other words, the value assignments $D_i = x_i$ are events in $\mathcal{F}$.

Note that, although this is not the standard way of modeling a probability space for embeddings, there is no formal obstacle to this interpretive shift. This construction allows linguistic expressions and their embeddings to be treated as events within a common probability space, which is necessary for studying the relationship between probabilistic quantities such as PMI and geometric operations in the embedding space.

3 Embedding Models: Assumptions

Once the probabilistic space has been established, we can characterize the distributional principle that connects contextual cooccurrence with proximity in the embedding space. Distributional embedding models are based on the general principle that the

¹Note that embeddings, like expressions, are not elementary disjoint events, but rather events in the σ -algebra $\mathcal{F}$.

representation of an expression is determined by the contexts with which it is associated.

In our probabilistic framework, the degree to which two expressions share contexts is captured by their joint probability $P(x, y)$ relative to their marginal probabilities $P(x)$ and $P(y)$. Thus, two expressions are strongly distributionally associated when they share more contextual probability mass than would be expected from their individual occurrence probabilities. This notion is naturally captured by Pointwise Mutual Information (PMI):

$$PMI(x, y) = \log \left(\frac{P(x, y)}{P(x)P(y)} \right),$$

which quantifies the degree to which the contextual events associated with x and y overlap beyond what would be expected if they were statistically independent. In terms of information content, with

$$I(x) = -\log P(x),$$

PMI can equivalently be written as

$$PMI(x, y) = I(x) + I(y) - I(x, y).$$

Our distributional assumption is that an embedding model reflects this contextual structure geometrically: expressions that exhibit a greater degree of contextual association should be represented as more similar according to some similarity function in the embedding space. Importantly, this assumption does not require a particular concrete definition of context. It only requires that the probabilistic space induced by the chosen notion of context captures the contextual associations that the embedding model is intended to represent.

To provide greater generality, we consider the Information Contrast Model (Amigó et al., 2020), a parameterized generalization of PMI:

$$ICM(x, y) = \alpha_1 I(x) + \alpha_2 I(y) - \beta I(x, y),$$

with the constraint

$$\alpha_1, \alpha_2 < \beta < \alpha_1 + \alpha_2.$$

The constraint $\alpha_1, \alpha_2 < \beta$ ensures that occurrence similarity to a non-informative expression

$$ICM(x, x_\top) = (\alpha_1 - \beta)I(x)$$

decreases as the expression becomes more specific, while the constraint $\beta < \alpha_1 + \alpha_2$ ensures that self-context similarity

$$ICM(x, x) = (\alpha_1 + \alpha_2 - \beta)I(x)$$

increases with expression specificity (Amigó et al., 2020). Therefore, our first assumption states that the contextual association between expressions is reflected by a similarity function over their embedding representations. Formally, assuming symmetry ($\alpha_1 = \alpha_2 = \alpha$) and setting $\alpha = 1$ (without loss of generality):

Assumption 1 (DISTRIBUTIONAL SEMANTICS)

There exists a continuous and differentiable function $\mathbf{sim} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ in the embedding space that reflects the degree of contextual association between two expressions:

$$\mathbf{sim}(\mathbf{v}_x, \mathbf{v}_y) = I(x) + I(y) - \beta I(x, y),$$

with $1 < \beta < 2$.

This assumption formalizes the distributional principle in terms of contextual joint occurrence rather than requiring an explicit similarity relation between individual contexts.

Note that the assumption prescribes neither the specific form of the function $\mathbf{sim}$ nor the concrete parameter β , beyond requiring that it lies within the specified range. In the case of SGNS and GloVe, for instance, (Ethayarajh et al., 2019) found a linear correlation between the Euclidean distance and $I(x) + I(y) - 2I(x, y)$.

Our second assumption is that embeddings are dense representations, in the sense that redundancy is minimized by maximizing independence across dimensions.

Assumption 2 (DIMENSIONAL INDEPENDENCE)

The embedding dimensions are statistically independent. For all $i = 1 \dots n$:

$$P(D_i = v_i, \dots, D_n = v_n) = \prod_{i=1 \dots n} P(D_i = v_i)$$

DIMENSIONAL INDEPENDENCE

This assumption is natural if embedding dimensions are interpreted as non redundant factors of variation. This idealization is also aligned with a common goal in representation learning, namely to reduce redundancy and obtain more informative coordinates. Although full independence is unlikely to hold exactly in practice, many studies show that increasing independence or reducing correlation across dimensions improves embedding quality (Mu et al., 2017; Ethayarajh, 2019; Yamagiwa et al., 2023; Opitz et al., 2020).

The third assumption is that the similarity function $\mathbf{sim}$ is rotationally invariant (i.e., rotation is an

isometry under the distance implied by sim). For instance, in (Yin and Shen, 2018), it is assumed that applying an orthogonal transformation (rotation or reflection) to the embedding space does not alter the semantic information as captured by similarity relations. This assumption is justified by the fact that coordinate axes may vary across training runs with equivalent semantics due to random initialization, numerical noise, or the choice of factorization basis.

Assumption 3 (ROTATIONAL INVARIANCE)

Similarity is unaffected by rotations of the embedding space:

$$R^T R = I \Rightarrow \text{sim}(\mathbf{v}_x, \mathbf{v}_y) = \text{sim}(R\mathbf{v}_x, R\mathbf{v}_y)$$

While the assumptions introduced in this section can be debated and may not be strictly satisfied by all embedding models in practice, they capture core principles that are widely invoked, either explicitly or implicitly, in distributional semantics. Rather than claiming that these assumptions hold universally, our goal is to make them explicit and to analyze their theoretical consequences. Under these conditions, we establish the precise setting in which the inner product satisfies the structural properties required of an operator that is linearly correlated with Pointwise Mutual Information. In this sense, the assumptions should be understood not as empirical claims about existing models, but as sufficient conditions that characterize when and why the inner product is a valid predictor of PMI in embedding spaces.

4 PMI Estimator

Grounding on the previous assumptions, we can establish the following lemma. Since sim is rotationally invariant and dimensions are statistically independent (see the proof² in Appendix A):

Lemma 1 *The similarity function sim has the form:*

$$\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = c_1(\|\mathbf{v}_x\|^2 + \|\mathbf{v}_y\|^2) + c_2\mathbf{v}_x \cdot \mathbf{v}_y + c_3.$$

²The proof for Lemma 1 shows that dimensional independence first lets similarity be written as a sum of coordinate-wise terms. Then, rotational invariance removes any dependence on the particular coordinate system: the function can only depend on quantities preserved by rotations. This forces the coordinate-wise term to have constant curvature, so it must be quadratic. Summing across dimensions leaves exactly the squared norms of both vectors, their dot product, and a constant.

From this lemma and the previous assumptions, we can derive the following theorem (see the proof³ in Appendix B):

Theorem 1 *Under our three main assumptions (DISTRIBUTIONAL SEMANTICS, DIMENSIONAL INDEPENDENCE and ROTATIONAL INVARIANCE) there exists three constants α^* , β^* and γ^* such that:*

$$\mathbf{v}_x \cdot \mathbf{v}_y = \alpha^*(I(x) + I(y)) - \beta^*I(x, y) + \gamma^*.$$

This theorem shows that the inner product of embeddings necessarily approximates a linear combination of the individual and joint information quantities, i.e., a generalization of PMI.

But moreover, it can be shown that, if the representation of the non informative representation $\mathbf{v}_\top$ lies at the origin, then the dot product is proportional to PMI. This additional condition is consistent with findings showing that terms carrying little information tend to lie close to the origin (Yokoi et al., 2020; Gao et al., 2019; Schakel and Wilson, 2015; Arefyev et al., 2018; Pagliardini et al., 2018).

Theorem 2 *Under our three main assumptions (DISTRIBUTIONAL SEMANTICS, DIMENSIONAL INDEPENDENCE and ROTATIONAL INVARIANCE), if $\|\mathbf{v}_\top\| = 0$ then:*

$$\mathbf{v}_x \cdot \mathbf{v}_y \propto \text{PMI}(x, y).$$

5 Practical Consequences

Assuming the inner product is proportional to PMI under the conditions of Theorem 2 has useful implications. For instance, we can estimate any probabilistic similarity measure of the form $\alpha_1 I(x) + \alpha_2 I(y) - \beta I(x, y)$ in terms of operations over vectors, as is the case of (Amigó et al., 2022):

$$\begin{aligned} \text{ICM}^V(\mathbf{v}_x, \mathbf{v}_y) &= (\alpha_1 - \beta)\|\mathbf{v}_x\|^2 \\ &+ (\alpha_2 - \beta)\|\mathbf{v}_y\|^2 + \beta\mathbf{v}_x \cdot \mathbf{v}_y \end{aligned}$$

since both are linearly correlated (See proof in Appendix C).

³The proof for Theorem 1 uses special cases, such as comparing an expression with the non-informative expression, comparing an expression with itself, and comparing orthogonal embeddings, to connect vector norms with information content. Then it substitutes those norm terms back into the general form from Lemma 1. This isolates the dot product and shows that it must be a linear combination of $I(x)$, $I(y)$, and $I(x, y)$.

We can also obtain an estimator $\widehat{\text{CondP}}(x, y)$ that is proportional to the logarithm of the conditional probability of expressions (See proof in Appendix D):

$$\widehat{\text{CondP}}(x, y) = \mathbf{v}_x \cdot \mathbf{v}_y - |\mathbf{v}_y|^2 \propto \log(P(x | y))$$

We can also derive an analogy function based on the definition in (Arora et al., 2016), where analogy is defined as a second order relationship. For instance, there is analogy between *man*, *woman*, *king* and *queen* because there exist a ratio correspondence between conditional probabilities of words x :

$$\frac{P(x|\text{queen})}{P(x|\text{king})} \simeq \frac{P(x|\text{woman})}{P(x|\text{man})}$$

Therefore, an analogy relationship holds among four embeddings $\vec{a}$, $\vec{b}$, $\vec{c}$, and $\vec{d}$ to the extent that, for any representation x in the embedding space, the following condition is satisfied:

$$\widehat{\text{CondP}}(\mathbf{v}_x, \vec{a}) - \widehat{\text{CondP}}(\mathbf{v}_x, \vec{b}) = \widehat{\text{CondP}}(\mathbf{v}_x, \vec{c}) - \widehat{\text{CondP}}(\mathbf{v}_x, \vec{d})$$

Interestingly, this implies that when the four embeddings have similar norms, the well-known algebraic relation by Mikolov holds, i.e., $\vec{d} = \vec{c} + \vec{b} - \vec{a}$ (see proof in Section E).

6 Conclusions

The theoretical result established in this article can be stated as follows: Assuming that (i) there exists a contextual space such that the embedding model, according to some similarity function, maps expressions that share contexts to similar representations; (ii) within this contextual space, the embedding dimensions are statistically independent; and (iii) the similarity function is rotationally invariant, then the inner product between embeddings is a linear combination of the individual and joint information content of the expressions. Furthermore, if the representation of the non-informative expression, i.e., the expression with maximum probability, lies at the origin of the embedding space, then the inner product directly approximates PMI over the corresponding contextual space.

This theoretical result has practical implications. Under the assumptions described above, we can straightforwardly derive embedding operators corresponding to semantic similarity, information content, semantic implication, and analogy.

Note that the significance of our result does not hinge on the universal validity of the assumptions used in our formal analysis. Some of these assumptions may be relaxed or violated by specific embedding models, particularly in practical or contextualized settings. However, our analysis clarifies that whenever distributional representations satisfy these minimal probabilistic and geometric conditions, the inner product is not merely a convenient choice, but the only operator capable of fulfilling the requirements imposed by Pointwise Mutual Information.

Limitations

This work is primarily theoretical and, as such, its conclusions are contingent on a set of explicit modeling assumptions. While these assumptions are motivated by common principles in distributional semantics, they may not be strictly satisfied by all embedding models used in practice.

First, the assumption of statistical independence across embedding dimensions is known to be violated in many pretrained representations. However there is empirical evidence that encouraging independence through postprocessing or regularization improves embedding quality. Our results therefore apply most directly to idealized or normalized embedding spaces that approximate this condition.

Second, the probabilistic interpretation of embeddings introduced in Section 2 is not conventional, as it models value assignments in embedding components as probabilistic events rather than random variables. Although this construction is formally valid and enables a unified probability space for expressions and embeddings, it may not align with alternative probabilistic or generative interpretations of representation learning.

Third, in our formulation, contexts are treated as elementary outcomes of the probability space, while expressions are represented as events defined over them. Consequently, the association between an expression and a particular context is necessarily binary: an expression either occurs in, or is associated with, a context or it does not. This discrete notion of expressions occurring in contexts is necessary in order to define PMI between expressions in its traditional form. A more general framework should allow this association to be gradual.

Finally, the paper does not include an empirical evaluation. While the goal of this work is to establish a uniqueness result rather than an empirical

estimator, experimental analysis could help assess how deviations from the assumptions affect the practical relationship between PMI and the inner product in modern embedding models. Exploring such robustness empirically is an immediate direction for future work.

Acknowledgements

This work has been funded by the European Union (NextGenerationEU) through the Recovery, Transformation and Resilience Plan, by the Ministry of Economic Affairs and Digital Transformation, and by the National Distance Education University (UNED) under cooperation agreement C039-21OT (ODESIA project); and by the Spanish Ministry of Science, Innovation and Universities (project ANNOTATE (PID2024-156022OB-C31), funded by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+); and TRAIL-LLMs project (AIA2025-163322-C65) in the framework of the coordinated project HumanAIze).

Use of AI assistants

Generative AI has been used for proofreading the manuscript and for proof-checking of the formal derivations.

References

- Enrique Amigó, Alejandro Ariza-Casabona, Victor Fresno, and M. Antònia Martí. 2022. [Information theory-based compositional distributional semantics](#). *Computational Linguistics*, 48(4):907–948.
- Enrique Amigó, Fernando Giner, Julio Gonzalo, and M. Verdejo. 2020. [On the foundations of similarity in information access](#). *Information Retrieval Journal*, 23:216–254.
- Nikolay Arefyev, Pavel Ermolaev, and Alexander Panchenko. 2018. [How much does a word weigh? weighting word embeddings for word sense induction](#). *ArXiv*, abs/1805.09209.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 EMNLP-IJCNLP*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. [Towards understanding linear word analogies](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.
- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. [Representation degeneration problem in training natural language generation models](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems 27*, pages 2177–2185.
- Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. [On the sentence embeddings from pre-trained language models](#). In *2020 Conference on EMNLP*, pages 9119–9130, Online. Association for Computational Linguistics.
- Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. [All-but-the-top: Simple and effective postprocessing for word representations](#). *CoRR*, abs/1702.01417.
- Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. 2020. [Deep metric learning with bier: Boosting independent embeddings robustly](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):276–290.
- Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. 2023. [Norm of word embedding encodes information gain](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2108–2130. Association for Computational Linguistics.
- Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. [Unsupervised learning of sentence embeddings using compositional n-gram features](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.
- Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. [Measuring word significance using distributed representations of words](#). *ArXiv*, abs/1508.02297.
- Hiroaki Yamaguchi, Momose Oyama, and Hidetoshi Shimodaira. 2023. [Discovering universal geometry in embeddings with ICA](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4647–4675, Singapore. Association for Computational Linguistics.

Zi Yin and Yuanyuan Shen. 2018. [On the dimensionality of word embedding](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. [Word rotator’s distance](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2944–2960, Online. Association for Computational Linguistics.

7 Appendix

A Proof for Lemma 1

In this appendix, we show that, under our three main assumptions (Probabilistic Similarity, Dimensional Independence, and Rotational Invariance) the similarity function must take the following form:

$$\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = c_1(\|\mathbf{v}_x\|^2 + \|\mathbf{v}_y\|^2) + c_2 \mathbf{v}_x \cdot \mathbf{v}_y + c_3.$$

Starting from Assumption 1, we have

$$\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = I(x) + I(y) - \beta I(x, y),$$

with $1 < \beta < 2$. Moreover, $I(x)$ can be expressed as $-\log P(x)$, where, as explained in Section 2,

$$P(x) = P(\mathbf{v}_x) = P(D_1 = x_1, \dots, D_n = x_n).$$

Given that the embedding dimensions are statistically independent (Assumption 2), we can decompose:

$$\begin{aligned} \text{sim}(\mathbf{v}_x, \mathbf{v}_y) &= \sum_i \alpha_1 I(D_i = x_i) \\ &\quad + \sum_i \alpha_2 I(D_i = y_i) \\ &\quad - \sum_i \beta I(D_i = x_i, D_i = y_i) \end{aligned}$$

Therefore, the summations may be expressed as a summation of functions:

$$\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = \sum_i g_i(x_i, y_i)$$

Considering that dimensional independence includes invariance under permutations of coordinates, then all g_i coincide up to additive constants, which can be unified into a global constant. Thus, we may write:

$$\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = \sum_{i=1}^n g(x_i, y_i)$$

for a single function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$.

Consider $\mathbf{v}_x = (u, 0, \dots, 0)$ and $\mathbf{v}_y = (v, 0, \dots, 0)$, $u, v \in \mathbb{R}$, then,

$$\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = g(u, v) + (n-1) \cdot g(0, 0),$$

where the second summation is constant and irrelevant for the shape of sim . Thus, in order to determine the shape of sim , the problem can be reduced to two real variables.

Consider $\mathbf{v}_x(\theta) = (u \cos \theta, u \sin \theta, 0, \dots, 0)$ and $\mathbf{v}_y(\theta) = (v \cos \theta, v \sin \theta, 0, \dots, 0)$, for any θ . By rotational invariance, $\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = \text{sim}(\mathbf{v}_x(\theta), \mathbf{v}_y(\theta))$, then using the separable form:

$$\begin{aligned} g(u, v) + g(0, 0) &= g(u \cos \theta, v \cos \theta) \\ &\quad + g(u \sin \theta, v \sin \theta) \quad \forall \theta. \end{aligned}$$

As the function $F(\theta) = g(u \cos \theta, v \cos \theta) + g(u \sin \theta, v \sin \theta)$ is constant, then $F''(\theta) = 0$. Computing the second derivative of F at 0, we obtain:

$$\begin{aligned} &u^2 \left(\frac{\partial^2 g}{\partial x^2}(u, v) - \frac{\partial^2 g}{\partial x^2}(0, 0) \right) \\ &+ 2uv \left(\frac{\partial^2 g}{\partial x \partial y}(u, v) - \frac{\partial^2 g}{\partial x \partial y}(0, 0) \right) \\ &+ v^2 \left(\frac{\partial^2 g}{\partial y^2}(u, v) - \frac{\partial^2 g}{\partial y^2}(0, 0) \right) = 0. \end{aligned}$$

Using the subindex notation for the partials, and setting $v = 0$, then $g_{x,x}(u, v) - g_{x,x}(0, 0) = 0$. Similarly, by setting $u = 0$, then $g_{y,y}(u, v) - g_{y,y}(0, 0) = 0$. Finally, by choosing (u, v) and $(u, -v)$, and subtracting the two equations, then $g_{x,y}(u, v) - g_{x,y}(0, 0) = 0$. As these identities are valid for every (u, v) , then the Hessian of g is constant. It implies that g must be a polynomial of degree at most two:

$$g(x, y) = c_1 x^2 + c'_1 y^2 + c_2 xy + c_3$$

for some constants c_1, c'_1, c_2, c_3 . In addition, the symmetry of the similarity, $\text{sim}(\mathbf{v}_x, \mathbf{v}_y) = \text{sim}(\mathbf{v}_y, \mathbf{v}_x)$, implies $g(x, y) = g(y, x)$, and therefore $c_1 = c'_1$.

Finally,

$$\begin{aligned} \text{sim}(\mathbf{v}_x, \mathbf{v}_y) &= \sum_{i=1}^n g(x_i, y_i) = \\ &c_1 \cdot \sum_i x_i^2 + c_1 \cdot \sum_i y_i^2 + c_2 \cdot \sum_i x_i y_i + n \cdot c_3 = \\ &c_1(\|\mathbf{v}_x\|^2 + \|\mathbf{v}_y\|^2) + c_2 \mathbf{v}_x \cdot \mathbf{v}_y + c'_3. \end{aligned}$$

B Proof for Theorems 1 and 2

B.1 Part I: Statements

First, we set four statements derived from Lemma 1 and assumptions 1 and 2 under different conditions of $\mathbf{v}_x$ and $\mathbf{v}_y$.

Being $\mathbf{v}_\top$ the embedding of the non informative expression $x_\top$ such that ($P(\mathbf{v}_\top) = 1$), then

$$I(x) + I(x_\top) - \beta I(x, x_\top) = (1 - \beta)I(x)$$

Then, according to Lemma 1 and assumption 1:

$$\begin{aligned} \text{sim}(\mathbf{v}_x, \mathbf{v}_\top) &= \\ &= c_1(\|\mathbf{v}_x\|^2 + \|\mathbf{v}_\top\|^2) + c_2 \mathbf{v}_x \cdot \mathbf{v}_\top + c_3 \\ &= (1 - \beta)I(x) \end{aligned} \quad (1)$$

In addition,

$$I(x_\top) + I(x_\top) - \beta I(x_\top, x_\top) = (1 - \beta)I(x_\top) = 0$$

Therefore:

$$\text{sim}(\mathbf{v}_\top, \mathbf{v}_\top) = (2c_1 + c_2)\|\mathbf{v}_\top\|^2 + c_3 = 0 \quad (2)$$

In addition, if $\mathbf{v}_x$ and $\mathbf{v}_y$ are orthogonal, then $\mathbf{v}_x \cdot \mathbf{v}_y = 0$ and assuming that dimensions are statistically independent, then

$$P(\mathbf{v}_x, \mathbf{v}_y) = P(\mathbf{v}_x)P(\mathbf{v}_y)$$

and

$$P(x, y) = P(x)P(y).$$

Therefore, $I(x, y) = I(x) + I(y)$ and:

$$\begin{aligned} \mathbf{v}_x \perp \mathbf{v}_y \Rightarrow \text{sim}(\mathbf{v}_x, \mathbf{v}_y) &= \\ &= c_1(\|\mathbf{v}_x\|^2 + \|\mathbf{v}_y\|^2) + c_3 \\ &= (1 - \beta)(I(x) + I(y)) \end{aligned} \quad (3)$$

Finally, according to Lemma 1 and assumption 1:

$$\begin{aligned} \text{sim}(\mathbf{v}_x, \mathbf{v}_x) &= \\ &= (2c_1 + c_2)\|\mathbf{v}_x\|^2 + c_3 \\ &= (2 - \beta)I(x) \end{aligned} \quad (4)$$

B.2 Part II: Theorem 1

Theorem 1 states that *under our three main assumptions (Probabilistic Similarity, Dimensional Independence and Rotational Invariance) there exists three constants α^* , β^* and γ^* such that:*

$$\mathbf{v}_x \cdot \mathbf{v}_y = \alpha^*(I(x) + I(y)) - \beta^*I(x, y) + \gamma^*.$$

On the basis of the previous statements, we can prove the theorem.

From Equation 4 we can derive:

$$\|\mathbf{v}_x\|^2 = \frac{2 - \beta}{2c_1 + c_2}I(x) - \frac{c_3}{2c_1 + c_2}$$

Replacing in Lemma 1 we obtain:

$$\begin{aligned} c_1 \left(\frac{2 - \beta}{2c_1 + c_2} (I(x) + I(y)) - \frac{2c_3}{2c_1 + c_2} \right) \\ + c_2 (\mathbf{v}_x \cdot \mathbf{v}_y) + c_3 = \\ = I(x) + I(y) - \beta I(x, y) \end{aligned}$$

Therefore:

$$\begin{aligned} \mathbf{v}_x \cdot \mathbf{v}_y &= \frac{1}{c_2} \left(1 - \frac{c_2(2 - \beta)}{2c_1 + c_2} \right) (I(x) + I(y)) \\ &\quad - \frac{\beta}{c_2} I(x, y) + \frac{c_3}{c_2} \left(1 + \frac{2c_1}{2c_1 + c_2} \right) \end{aligned}$$

Therefore, Theorem 1 is proved for the specific values:

$$\begin{aligned} \alpha^* &= \frac{1}{c_2} \left(1 - \frac{c_2(2 - \beta)}{2c_1 + c_2} \right) \\ \beta^* &= \frac{\beta}{c_2} \\ \gamma^* &= \frac{c_3}{c_2} \left(1 + \frac{2c_1}{2c_1 + c_2} \right) \end{aligned}$$

B.3 Part III: Theorem 2

Theorem 2 states that *under our three main assumptions (Probabilistic Similarity, Dimensional Independence and Rotational Invariance), if $\|\mathbf{v}_\emptyset\| = 0$ then:*

$$\mathbf{v}_x \cdot \mathbf{v}_y \propto PMI(x, y).$$

If $\|\mathbf{v}_\top\| = 0$ then, according to Equation 2, we obtain $c_3 = 0$. Then, from Equation 4 we obtain:

$$\|\mathbf{v}_x\|^2 = \frac{2 - \beta}{2c_1 + c_2}I(x)$$

Substituting into Equation 3 for the orthogonal case, i.e. $\mathbf{v}_x \perp \mathbf{v}_y$ with $c_3 = 0$:

$$\begin{aligned} \frac{c_1(2 - \beta)}{2c_1 + c_2} (I(x) + I(y)) \\ = (1 - \beta) (I(x) + I(y)) \end{aligned}$$

and therefore:

$$\beta = 1 - \frac{c_1(2 - \beta)}{2c_1 + c_2}$$

Therefore, $\alpha^* = \beta^*$ and $\gamma^* = 0$. Then:

$$\begin{aligned} \mathbf{v}_x \cdot \mathbf{v}_y &= \alpha^* (I(x) + I(y)) - \alpha^* I(x, y) \\ &\propto I(x) + I(y) - I(x, y) = \text{PMI}(x, y) \end{aligned}$$

C Formal Proof for the Similarity Operator

In this appendix we formally prove that the generalized Information Contrast Model similarity:

$$\text{ICM}(x, y) = \alpha_1 I(x) + \alpha_2 I(y) - \beta I(x, y)$$

and the expression described in (Amigó et al., 2022):

$$\begin{aligned} \widehat{\text{ICM}}(\mathbf{v}_x, \mathbf{v}_y) &= (\alpha_1 - \beta) \|\mathbf{v}_x\|^2 \\ &\quad + (\alpha_2 - \beta) \|\mathbf{v}_y\|^2 \\ &\quad + \beta \|\mathbf{v}_x\| \|\mathbf{v}_y\| \cos(\vec{v}_1, \vec{v}_2) \end{aligned}$$

are linearly correlated

From Theorem 2, we can derive:

$$\begin{aligned} \text{PMI}(x, y) &= a(\mathbf{v}_x \cdot \mathbf{v}_y) \\ I(x) &= \text{PMI}(x, x) = a(\mathbf{v}_x \cdot \mathbf{v}_x) = a\|\mathbf{v}_x\|^2 \end{aligned}$$

Therefore:

$$\begin{aligned} \alpha_1 I(x) + \alpha_2 I(y) - \beta I(x, y) &= \\ &= \alpha_1 (a\|\mathbf{v}_x\|^2) + \alpha_2 (a\|\mathbf{v}_y\|^2) \\ &\quad - \beta (a(\mathbf{v}_x \cdot \mathbf{v}_y)) \\ &= \alpha_1 a\|\mathbf{v}_x\|^2 + \alpha_2 a\|\mathbf{v}_y\|^2 \\ &\quad - \beta a \|\mathbf{v}_x\| \|\mathbf{v}_y\| \cos(\mathbf{v}_x, \mathbf{v}_y) \\ &= a(\alpha_1 \|\mathbf{v}_x\|^2 + \alpha_2 \|\mathbf{v}_y\|^2 \\ &\quad - \beta a \|\mathbf{v}_x\| \|\mathbf{v}_y\| \cos(\mathbf{v}_x, \mathbf{v}_y)) \end{aligned}$$

and therefore:

$$\begin{aligned} \alpha_1 I(x) + \alpha_2 I(y) - \beta I(x, y) \\ \propto \alpha_1 \|\mathbf{v}_x\|^2 + \alpha_2 \|\mathbf{v}_y\|^2 - \beta \|\mathbf{v}_x\| \|\mathbf{v}_y\| \cos(\mathbf{v}_x, \mathbf{v}_y) \end{aligned}$$

D Formal Proof for the Conditional Probability Operator

In this appendix we formally prove that the conditional probability operator is proportional to the

conditional probability of the expressions that represent:

$$\widehat{\text{CondP}}(x, y) = \mathbf{v}_x \cdot \mathbf{v}_y - \|\mathbf{v}_y\|^2 \propto P(x|y)$$

Assuming that $\mathbf{v}_x \cdot \mathbf{v}_x \propto \text{PMI}(x, y)$, we can expand $\log(P(x|y))$ as follows:

$$\begin{aligned} \log(P(x|y)) &= -\log\left(\frac{P(y)}{P(x, y)}\right) \\ &= -\log\left(\frac{P(x)P(y)}{P(x, y)}\right) + \log(P(x)) \\ &= \text{PMI}(x, y) - I(x) \\ &= a((\mathbf{v}_x \cdot \mathbf{v}_y) - \|\mathbf{v}_x\|^2) \\ &\propto (\mathbf{v}_x \cdot \mathbf{v}_y) - \|\mathbf{v}_x\|^2 \\ &= \text{CondP}(x, y) \end{aligned}$$

E Appendix. The Derivation of Mikolov Embedding Algebra

In this appendix, we formally prove that when the four embeddings have similar magnitudes, the well-known algebraic relation by Mikolov holds.

The expression:

$$\begin{aligned} \widehat{\text{CondP}}(\mathbf{v}_x, \vec{a}) - \widehat{\text{CondP}}(\mathbf{v}_x, \vec{b}) \\ = \widehat{\text{CondP}}(\mathbf{v}_x, \vec{c}) - \widehat{\text{CondP}}(\mathbf{v}_x, \vec{d}) \end{aligned}$$

can be written as:

$$\mathbf{v}_x \cdot \vec{a} - |\vec{a}|^2 - \mathbf{v}_x \cdot \vec{b} + |\vec{b}|^2 = \mathbf{v}_x \cdot \vec{c} - |\vec{c}|^2 - \mathbf{v}_x \cdot \vec{d} + |\vec{d}|^2$$

If the four embeddings have similar length:

$$\mathbf{v}_x \cdot \vec{a} - \mathbf{v}_x \cdot \vec{b} = \mathbf{v}_x \cdot \vec{c} - \mathbf{v}_x \cdot \vec{d}$$

Therefore:

$$\sum_i (a_i - b_i - c_i + d_i) x_i = 0$$

This is true only if $\forall_{i=1 \dots n} a_i - b_i - c_i + d_i = 0$, that is, $\vec{d} = \vec{c} + \vec{b} - \vec{a}$.