AIMS Mathematics, 10(2): 3961–3982.
DOI: 10.3934/math.2025185
Received: 12 November 2024
Revised: 02 February 2025
Accepted: 18 February 2025
Published: 27 February 2025

http://www.aimspress.com/journal/Math

Research article

Mathematical features of semantic projections and word embeddings for automatic linguistic analysis

Pedro Fernández de Córdoba, Carlos A. Reyes Pérez and Enrique A. Sánchez Pérez*

Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

* **Correspondence:** Email: easancpe@mat.upv.es; Tel: 0034963877000; Fax: 00343877669.

Abstract: Embeddings in normed spaces are a widely used tool in automatic linguistic analysis, as they help model semantic structures. They map words, phrases, or even entire sentences into vectors within a high-dimensional space, where the geometric proximity of vectors corresponds to the semantic similarity between the corresponding terms. This allows systems to perform various tasks like word analogy, similarity comparison, and clustering. However, the proximity of two points in such embeddings merely reflects metric similarity, which could fail to capture specific features relevant to a particular comparison, such as the price when comparing two cars or the size of different dog breeds. These specific features are typically modeled as linear functionals acting on the vectors of the normed space representing the terms, sometimes referred to as semantic projections. These functionals project the high-dimensional vectors onto lower-dimensional spaces that highlight particular attributes, such as the price, age, or brand. However, this approach may not always be ideal, as the assumption of linearity imposes a significant constraint. Many real-world relationships are nonlinear, and imposing linearity could overlook important non-linear interactions between features. This limitation has motivated research into non-linear embeddings and alternative models that can better capture the complex and multifaceted nature of semantic relationships, offering a more flexible and accurate representation of meaning in natural language processing.

Keywords: Lipschitz function; semantic projection; word embedding; model; Arens-Eells
Mathematics Subject Classification: 51F30, 68Q55

1. Introduction

One of the main conceptual tools of natural language processing (NLP) is the identification of sets of words with a subset of a vector space, called a word embedding [6]. Although the boundaries of this identification are often unclear to users, the main feature of the vector space used in this

representation is intended to be the norm, so that the distance that words inherit from their representation provides a measure of the semantic similarity between them ([16, 17]). The so-called cosine similarity (computed by using the scalar product) is another common tool to achieve this goal. Sometimes, other characteristic elements of normed spaces—linear structure, dual space functionals [14]—are also included in the use of the representations and take some role in the model. For example, linear functionals can be used as indicators of the level at which a certain property is shared by the words considered (see [10], where a mean of these functionals is called semantic projection).

All these features of linear space representation are intended to be parallel linear operations and relations that can be established between words in a semantic environment [9]. For example, to some extent, addition, subtraction, and even scalar multiplication could have a meaning as relations between semantic elements when interpreted in the model [15].

However, linear space structure is rather rigid, and often the operations defined there do not correspond to the relations that are needed to be established in a model for semantic environments. Of course, this circumstance is perfectly known by the scientists that work on the topic, and the models are adapted to overcome this obstacle. For example, the larger the dimension of the linear space, the greater the chance of obtaining a good representation, since more features of the words can be represented. The co-occurrence of words in similar linguistic contexts can be a measure of proximity, so the vectors representing similar words have to be similar as well, i.e. the distance (measured by the norm) between them has to be small.

In this paper, we focus on the mathematical environment provided by the model based on what are called semantic projections, closely related to the framework of distributional semantics and word embeddings (see, for example, [5, 13] and references therein). In [10], an intuitive description of a semantic projection is given. Essentially, given a vector space representation of a semantic corpus E provided by a word embedding, a semantic projection associated with a certain property of a given subset $S \subseteq E$ is a real function $p : S \to \mathbb{R}$ that scales (by means of a real number) how the property affects each of the elements of S. For example, if the subset S of E under consideration is that of "animals", an example of semantic projection would be given by the notion of "size". In the context of mathematical analysis, the representation of such a function, as defined in [10], is exactly an element of the dual space E^* of E (perhaps after a translation to get $p(0) = 0$.) With some slightly different meanings, the same fundamental idea ("semantic projections can be represented as linear transformations") can be found in many papers (see, for example, [4, 8]; also, refer to the references in [10]). It should be noted that closely related concepts are also applied in non-linguistic contexts, such as image processing [8, 19].

While this might give a good and simple approximation to the idea trying to be modeled (the theoretical concept of semantic projection), it lacks some fundamental inadequacies that can be easily solved by enriching the mathematical tools used. As already said, the main problem is that linearity, both of the vector space and the function providing the semantic projection, is often not the best assumption to be made.

To summarize, we are interested in analyzing to what extent the semantic projection model is constrained by the intrinsic properties of the linear space structure, as well as proposing a solution to the problem of how these constraints can be overcome. Thus, our aim is to give a general idea of what could be the most general correct framework for considering an unambiguous vector representation

model. From the theoretical point of view, the natural setting for fixing the relations between distances and norms can be found in the theory of so-called Arens-Eells spaces (also known as Lipschitz-free spaces). In our opinion, this should be the correct starting point for a vector space representation of a semantic word model. In this context, the natural way to define the semantic projection is by a real Lipschitz function, rather than by a linear functional. This basic difference completely changes the method of model building for semantic environments, providing a new way of understanding the interaction of semantic embeddings and newly introduced terms in conceptual environments, supported by the mathematical development presented here.

2. Mathematical concepts and basic properties of word embeddings

Let us consider the mathematical aspects of what a word embedding is. Essentially, it consists of assign a vector of a linear space to each word that we consider in the original set we want to represent. When modeling a word embedding of this type, one of the normal objectives is to reduce the dimension of the vector space in which a set of words can be "semantically embedded". Consider a set of words W with a rich network of semantic interactions operating among them. There are several methods to compute the coordinates of the words in the vector space that is chosen; a common procedure consists of using the distribution of co-occurrences in a corpus of documents (distributional semantic model [13]). For example, and simplifying the model, if a list of features is given, a vector in which each coordinate gives a measure of how much a single word satisfies a given property of the list, provides a word embedding.

In such context, if the cardinality of the set is $|W| = n$, we have that each element w of the set W can be identified, for example, with the vector e_w of the canonical basis $\{e_w : w \in W\}$ of $\mathbb{R}^n$. The distance among any two different words in W is then given by a constant ($\sqrt{2}$ for the Euclidean norm), but it can be modulated by considering convenient weights. However, it has to be taken into account that the use of a *norm* for defining a distance that emulates the semantic similarity among words is already a huge restriction. *A priori*, not all distances can be written using such an embedding if the identification is made as explained and we force to consider in $\mathbb{R}^n$ the Euclidean norm. Also, the linear structure of the space is intended to have a meaning in the semantic model; we will come back to this point later on.

But the intrinsic goal of a word embedding is dimensionality reduction. That is, we have to find a natural number $1 \leq r \leq n$ such that there is a map $i : W \to \mathbb{R}^r$ satisfying that the distance d provided by the Euclidean norm on $\mathbb{R}^r$—or other norm in the space—represents the "semantic distance" among words.

From the point of view of the theory of Banach spaces, this goal—which is natural in applied contexts such as the NLP cited above—is a bit confusing. Readers familiar with the geometric theory of Banach spaces know that the use of norms when taking vector representations is a methodological choice that sometimes gives problems, due to the rigidity of some features of the norms. These features are, in fact, the very ones that make this theory so rich. However, these problems can be avoided if an adequate theoretical context—even in the domain of Banach spaces—is chosen. Furthermore, and primarily, normed spaces have a linear structure in addition to the metric one, which provides more tools for dealing with computational models.

The best option for improving the adaptability of the models, beyond normed spaces, is the more

general environment of metric spaces, which provide the context in which we work in the present article. Of course, linearity is lost, but we will see that we can recover it in some sense in what follows. Although more complex options are often considered, the class of mathematical relations that can help for describing semantic environments is not so big. Complexity of the models, although of course makes them richer, often adds non-useful or even inexact relations that could lead to confusion. On the other hand, a distance between terms seems to be easy to understand for experts in semantics: the notion of how "close" are two words regarding their meanings is intuitive and easy to handle. Since sometimes we would need more instruments to introduce other properties, we will use the informal notion of "enriched" metric space which could include more mathematical features as linear operations, other metrics, or equivalence relations (see [3]).

We use standard analytical [1, 7] and topological concepts [12]. If $(E, \|\cdot\|)$ is a normed space, we write E^* for its dual space, that is the Banach space of all linear continuous real valued functions $x^* : E \to \mathbb{R}$ with the supremum norm. Let (X, d) be a metric space in which we fix a singular point; we write 0 for it. (X, d) is then said to be a pointed metric space. If (Y, ρ) is another metric space, we say that a map $f : X \to Y$ is Lipschitz if there is a constant $K > 0$ such that

$$\rho(f(x), f(y)) \leq K\, d(x, y), \quad x, y \in X.$$

The least such constant K is called the Lipschitz norm of f, written as $Lip(f)$. Fix the metric space (Y, ρ) to be a normed space $(E, \|\cdot\|)$. Let E be a Banach space. The space $Lip_0(X, E)$ of Lipschitz maps such that $f(0) = 0$ is the Banach space of Lipschitz operators from X to E that vanish at 0 with the Lipschitz norm $Lip\,(.)$. The reader can find all the information needed on these spaces in [7]. For the particular case $E = \mathbb{R}$, we write $X^{\#} = Lip_0(X) = Lip_0(X, \mathbb{R})$, which is called, by abuse of notation, the Lipschitz dual of X. The space $Lip_0(X)$ has what is known as a predual, called the Arens-Eells space $Æ(X)$ [2] that will play a central role in the present paper. Thus, we have that $Æ(X)^* = Lip_0(X) = X^{\#}$. Sometimes is also called the free space associated to X.

Let us describe first the elements of this space. A molecule on X is a real function $m : X \to \mathbb{R}$ on X of finite support for which $\sum_{x \in X} m\,(x) = 0$. We write $\mathcal{M}(X)$ for the real linear space of all molecules on X. For $x, x' \in X$, the molecule $m_{xx'}$ is defined by $m_{xx'} = \chi_{\{x\}} - \chi_{\{x'\}}$. Here, χ_S is the characteristic function of the set S, which equals 1 in x if $x \in S$ and 0 if $x \notin S$.

If $m \in \mathcal{M}(X)$, we write $m = \sum_{j=1}^{n} \lambda_j m_{x_j x'_j}$, and we can define a norm for the space by

$$\|m\|_{\mathcal{M}(X)} := \inf \left\{ \sum_{j=1}^{n} \left|\lambda_j\right| d(x_j, x'_j),\ m = \sum_{j=1}^{n} \lambda_j m_{x_j x'_j} \right\},$$

where the infimum is taken over all representations of the molecule m; note that the function that we define as a molecule allows to be written in different ways, with different elements, so it is not uniquely defined. This forces to consider all possible representations.

The completion of the normed space $(\mathcal{M}(X), \|.\|_{M(X)})$ is exactly the space $Æ(X)$. It can be easily seen that the operator $\iota : X \to Æ(X)$ given by $\iota(x) = m_{x0}$ embeds X in $Æ(X)$ isometrically.

A relevant result of the theory of Lipschitz operators gives that, for every $T \in Lip_0(X, E)$, there exists a unique linear map $T_L : Æ(X) \longrightarrow E$ such that $T = T_L \circ \iota$. Then the one-to-one relation $T \longleftrightarrow T_L$ establishes an isomorphism between the vector spaces $Lip_0(X, E)$ and the space of linear

operators $\mathcal{L}(Æ(X), E)$. This, in fact, proves that the spaces $X^{\#}$ and $Æ(X)^*$ are isometrically isomorphic, as we said above, via the linearization $R(f) := f_L$, where $f_L(m) = \sum_{x \in X} f(x)m(x)$. The map T_L is often called the linearization of T, [18, Theorem 2.2.4 (b)]. The interested reader can find more information on these spaces and operators in [7, 11].

3. Semantic environment and semantic projections

As we explained in the introduction, we base our work on the idea of semantic projection. This notion has been used in several recent research works, sometimes with rather different meanings. In general, this notion reveals a conceptual relationship between a term (a linguistic element endowed with a certain semantic value) and a semantic environment (a structured union of conceptual terms defining a meaningful context). The present paper does not go into the substantiation of these notions, which obviously raise deep semiotic questions. Instead, we try to build a formal tool useful for automatic language analysis, so we will focus on formal aspects. This means that, in what follows, a semantic term is a (single, part of, union of) word(s), and a semantic environment is a set of terms endowed with a metric with (perhaps) some relations between them. A semantic projection will simply be a real-valued function with some weak requirements. We will write P_w for the semantic projection $P_{\{w\}}$.

Definition. Let w be a semantic term, and consider a semantic environment S. We say that $P_S(w) \in \mathbb{R}$ is a semantic projection of w onto S if it gives a quantification of how the term w is represented by the meaning provided by S, in such a way that $|P_S(w)| \leq |P_w(w)|$.

In developing the model, the semantic projections are assumed to be monotone. That is, if S is fixed, we can compute the projection onto S of two terms w_1 and w_2, and $P_S(w_1) \leq P_S(w_2)$ if the meaning of w_1 is worse represented in S than that of w_2. For example, if S is defined by a single term, w_2 is a synonym of this term, and w_1 is an antonym, a reasonable semantic projection must satisfy $P_S(w_1) \leq 0 \leq P_S(w_2)$. However, this requirement cannot be strictly formalized, since the notion of "being worse represented", or even "being an antonym" is often vague.

Consequently, we will focus on mathematical definitions. The main reference we consider for the notion of semantic projection comes from the paper by Grand et al ([10]). In this paper, the essential idea (which we strongly follow in the present work), is that, whenever we have a semantic embedding in a vector space, we can use its linear structure to make this model provide information about features of the elements represented by the embedding. The way to do this is by defining a linear functional that fits the elements of the embedding by a line provided by a synonym-antonym pairs. As long as the elements of the word embedding are well represented in the vector space and the feature to be analyzed fits the way they are represented, we will obtain semantic information without the need to increase the terms of the word embedding.

Thus, the mathematical object obtained is a linear functional of the dual of the vector space in which the semantic embedding is located. As stated in [10], this method allows estimating human knowledge about semantic categories with limited effort in terms of human supervision, and preserves the idea that Euclidean space operations can be used to model semantic relations (see e.g. [4, 16]). Moreover, it is not necessary to modify the word embedding to enrich the features of the model in terms of representability, since the new semantic items are introduced as linear functionals belonging to the dual space, and not to the original space, which remains unchanged.

In this paper, we assume these general assumptions with one fundamental change: we avoid focusing the effectiveness of the model on the Euclidean structure of the vector representation, considering instead only a metric space. The linear functionals are then exchanged for Lipschitz functions, providing a more flexible method for introducing the new semantic element into the model. Thus, the dual E^* of the Euclidean space $(E, \|\cdot\|)$ to which the linear functionals belong is replaced by the Lipschitz dual $Lip(X)$ of the metric space (X, d), which is, in our case, the substrate of the model.

4. General proposed model

After the contextualization of the main models used to consider a word embedding, in this section we will discuss what would be the most general way to approach the problem of mathematical representations of semantic environments. Far from being the obvious best solution, representations in normed space have been widely used, as we said in the previous section. It is then natural to try to use the algebraic properties of vector spaces (i.e., linearity) as a complementary tool for these representations. Often, this only leads to problems since it is complicated to give a "semantic role" to vector addition and multiplication by scalars. Therefore, it seems more convenient to start by giving a model centered on metric properties and, if necessary, add some other algebraic or topological properties.

In this context, associating a vector to each word is a simple way to encode the information we have about it. Distances to other words can be defined in another way, since we do not need to use the natural metric associated to a norm. For example, if we have a list of "semantically independent words", like

$$M = \{\text{"}door\text{"}, \text{"}stone\text{"}, \text{"}lake\text{"}, \text{"}airplane\text{"}\},$$

we can endow it with the discrete metric, so that the distance between all of them is equal to one. For this we do not need any vector representation.

Remark. An additional tool that could be used for semantic mathematical models would be the use of fuzzy metrics instead of metrics. According to the original works by Zadé [20], fuzzy mathematics provides useful tools for NLP. This could relax strict metric relationships, which are sometimes not easy to fix. On the other hand, the use of fuzzy set theory could also be useful for semantic models, since often the membership of one element to the semantic neighbor of another element is not absolute, but can be estimated in probabilistic terms.

4.1. The mathematical construction of the model

As explained above, we chose the option of working with metric ideas. Consequently, the most elementary model is a metric space, to which we can add some additional structure. For example, in the case we are concerned with in this paper, we consider a subspace that has linear structure.

Let $(E, \|\cdot\|)$ be a normed space, and consider a (finite) set of semantic items W (i.e., words) and a semantic embedding $i : W \hookrightarrow E$. In this setting, consider a subspace L of E containing the representations for which we want to preserve the linear structure (under the hypothesis that this linear structure has a meaning in the model). Since we are working in a finite dimensional setting, we can write E as a direct sum of two subspaces, let us say L and V. Thus, $E = L \oplus V$, where the $1-$norm is taken in the sum, that is

$$\left\|(x, y)\right\| = \|x\| + \|y\|, \quad (x, y) \in L \oplus V.$$

Note that, in fact, we have two semantic embeddings $i_L : W_L \to L$ and $i_V : W_V \to V$. That is, the (global) semantic embedding is defined from a disjoint union of subsets (by definition, i is assumed to be injective), $W = W_L \cup W_V$ into E as

$$i = i_L \otimes i_V : W = W_L \cup W_V \hookrightarrow L \oplus V = E,$$

where $i_L \otimes i_V(x) = (i_L(x), 0)$ for $x \in W_L$ and $i_V \otimes i_V(y) = (0, i_V(y))$ for $y \in W_V$. On the other hand, we can consider different metrics than the one provided by a norm in E, at least for the second part V, while we preserve the norm for L. Let us endow V with a generic metric, d_V, and consider the restrictions to L and V, $\|\cdot\|$ and d_V, respectively.

Remark. However, note that for the model, the only elements of $V$with semantic meaning are the ones provided by the original elements of the set. That is, only $i_V(W_V)$ has to be considered for the model. We can introduce the rest of the elements of V in it, but since the linear structure of V does not have any meaning, we can remove it from the model, letting just $i_V(W_V) = S \subseteq V$ instead.

Let us explain how the properties concerning the semantic structure are introduced in the model. Let us recall that these properties are modeled are by means of real functions defined on the vector representation E. We call to such a function an index (we write I).Using these indices, we can evaluate the level of a given property, or separate disjoint sets that satisfy a given property. Let us see how mathematics is used to do this. We write the model for all V instead of the specific subset $i_V(W_V)$, but note that the construction makes sense for any subset $i_V(W_V) \subset S \subset V$.

- Normally, the linear structure is requested to be preserved—if it makes sense in the model—at least in the subspace L of the linear support space E. Thus, the type of function referred to above is represented by a continuous linear functional k of the dual of the linear space. If we assume that only in the subspace L does the linear structure make sense, the functional k belongs to L^*.
- However, a real function that is only compatible with the metric structure of the space is a real-valued Lipschitz function and not a linear functional. Let us recall that a real Lipschitz function φ can be considered as an element of the dual of the Arens-Eells space $Æ(V)$, that is $(Æ(V))^* = Lip(V)$, (where V is considered only as a metric space), as we explained in the Introduction.
- Since the vector support of the model is finally written as a direct sum $L \oplus V$ of the Banach space $(L, \|\cdot\|)$ and the metric space (V, d), any index required for the model is written as the map

$$k \oplus \varphi : L \oplus V \to \mathbb{R},$$

that acts on any element $x = (x_L, x_V) \in L \oplus V$ as

$$k \oplus \varphi(x) = k \oplus \varphi((x_L, x_V)) = k(x_L) + \varphi(x_V).$$

Since both maps are, in particular, Lipschitz (any continuous linear map is Lipschitz too), we have

$$\begin{aligned} \left|k \oplus \varphi(x) - k \oplus \varphi(y)\right| &= \left|k \oplus \varphi((x_L, x_V)) - k \oplus \varphi((y_L, y_V))\right| \\ &= \left|k(x_L) - k(y_L)\right| + \left|\varphi(x_V) - \varphi(y_V)\right| \\ &\leq \max\{\|k\|, Lip(\varphi)\} \cdot (\|x_L - y_V\| + d(x_V, y_V)) \end{aligned}$$

for each pair of elements $x = (x_L, x_V)$ and $y = (y_L, y_V)$.

Summarizing the ideas we have presented above, the model is based on a semantic embedding of W into the direct sum $L \times Æ(V)$. The main feature of this representation is that it increases the dimension of the space, since each element m_{x0}, $x \in W_V$, is linearly independent of the rest of functions as m_{y0}, $y \in W_V \setminus \{x\}$. Roughly speaking, while the elements represented in L preserve the linear structure, the elements in W_V remain free from the algebraic point of view. Consequently, any index (real Lipschitz function) I that can be used to determine any property of the terms of W will be represented as an addition of a linear functional k and a Lipschitz function φ. This avoids the problem of having an over-structured model in which the index is assumed to be a linear functional. The index acts on the nonlinear part of the model as a Lipschitz function, which preserves the main property of the model, i.e., the metric. In the context of semantic projections, we identify the indexes presented here with a real-valued semantic projection, as done in [10] for the linearity-preserving case.

The following factorization diagram represents the construction represented above.

$$\begin{array}{ccc} W = W_L \cup W_V & \xrightarrow{I} & \mathbb{R} \\ {\scriptstyle i_L \otimes i_V}\downarrow & & \uparrow{\scriptstyle [i]} \\ L \oplus V & \xrightarrow{k \oplus \varphi} & \mathbb{R} \end{array},$$

where k belongs to the topological dual of L^*, φ is a Lipschitz function in $Lip(V)$, and the identity map $[i]$ can be changed by a multiplication in case we want to modify the scale of the function I.

Finally, note that this scheme has to be adapted for each index (semantic projection in the sense of [10]). Therefore, it may happen that the subspaces L and V change in each case, or even that one of them is trivial (i.e., equal to $\{0\}$) for some semantic projections).

4.2. Semantic environments and semantic projections with an easy example

Let W be a semantic environment and consider a semantic embedding into a vector space E. Assume that we want to quantify a given semantic property of W as a semantic projection. According to [10], it can be done by using a *linear* functional f of E^*, in case linearity plays any role, as is usually assumed ([9, 17]). It is commonly accepted that the model of W provided by E uses three different features of the vector (normed) space: the Euclidean norm, the scalar product (that provides the cosine "distance"), and the linear structure. The properties of the functional f (it is linear and satisfies the Lipschitz inequality, since it is continuous) fit the given underlying structure of E.

In general, the justification of the (rather restrictive) linearity assumption comes from the fact that linear structure models terms. Let us explain this with an example. Suppose that we are considering multi-word terms in a semantic environment that contains the words $\{cow, dog, cat, big, small\}$. We can consider the terms *big dog* and *small dog*. The following equations make sense: $big\,dog - dog = big$, $small\,dog = small + dog$, and so we can infer that $(big\,dog) - (small\,dog) = big - small$, which represents, according to the compositional principle, that the relation among *big dog* and *small dog* is the same as that between *big* and *small*. The norm of the difference of the vectors representing the words *big* and *small* is the same as the norm of the difference of *big dog* and *small dog*.

Of course, we can extend the vector addition (and even the multiplication by scalars) to all the linear space E that acts as a model of W. The problem is that often, this extension does not work for all the terms, although the fundamental relation given by the metric—the norm in this example— is still

preserving its meaning in the model. For example, the term $cat - dog$ does not have a clear meaning in the model.

Let us perform a (non-linear) semantic projection ψ representing the notion of "size" in the model. Let us write W for the set of all compositions of different sizes of cows, dogs and cats. We follow the next steps.

- Consider the (trivial) word embedding $i : W \hookrightarrow E = \mathbb{R}^5$ given by $i(cow) = (1, 0, 0, 0, 0)$, $i(dog) = (0, 1, 0, 0, 0)$, $i(cat) = (0, 0, 1, 0, 0)$, $i(big) = (0, 0, 0, 1, 0)$, and $i(small) = (0, 0, 0, 0, 1)$.
- Take the linear space L to be the linear span

$$L = span\{(0, 0, 0, 1, 0), (0, 0, 0, 0, 1)\}.$$

- The distance for the complement subspace

$$V = span\{(1, 0, 0, 0, 0), (0, 1, 0, 0, 0), (0, 0, 1, 0, 0)\},$$

  (which has to be considered as a metric space), is also given by the Euclidean norm, that is $d(v, w) = \|v - w\|_2$.
- The semantic projection is defined as follows:
  1) for the linear part, put $f(a(0, 0, 0, 1, 0) - b(0, 0, 0, 0, 1)) = a - b$.
  2) For the non-linear (Lipschitz) part, define $\varphi((1, 0, 0, 0, 0))) = 20$, $\varphi((0, 1, 0, 0, 0)) = 2$, $\varphi((0, 0, 1, 0, 0)) = 1$ and $\varphi((0, 0, 0, 0, 0)) = 0$. If we write S for the subset of V given by these four elements, we clearly have that $Lip(\varphi|_S) = 20$. These specific numerical assignments serve as demonstrative values to help readers understand the mathematical concept being presented.
  Define now φ for the rest of the elements of V as the average of the McShane-Whitney extensions of φ, that is

$$\varphi(v) = \frac{1}{2} \sup_{s \in S} \Big(\varphi(s) - 20 \cdot d(s, v)\Big) + \frac{1}{2} \inf_{s \in S} \Big(\varphi(s) + 20 \cdot d(s, v)\Big), \quad v \in V.$$

  By the McShane-Whitney theorem this is a Lipschitz map belonging to $V^{\#} = Lip(V) = Æ(V)^*$ with Lipschitz norm $Lip(\varphi) = 20$.
  However, note that the elements of V that are not in S have no meaning in the model. Roughly speaking, the formula "cow + big dog" cannot be considered as a meaningful element of the word embedding if no explicit explanation is provided. If this is done, we would be extending the linearity assumptions on how the model works. This would force to use the restricted model given by

$$S \cup L \hookrightarrow Æ(S) \oplus L,$$

  and so any semantic projection would be written as $\phi = \varphi \oplus f \in Lip(S) \times L^*$.

The result is a semantic projection ψ that represents adequately the notion of "size" in the model. Indeed, this function makes sense for the set $S \times L$, but it can be computed for all the elements of $E = \mathbb{R}^5$.

Let us compute some concrete situations.

(a) The notion "big cat" would be given by the representation $(0, 0, 1, 1, 0) \in E$. Then, its quantification is given by $\psi(i(big\, cat)) = \varphi((0, 0, 1, 0, 0)) + f((0, 0, 0, 1, 0)) = 1 + 1 = 2$. The composed term "small cat" would be evaluated by $\psi((0, 0, 1, 0, -1)) = 1 - 1 = 0$.

(b) The size of the terms "cow" and "dog" can also compared with ψ. Indeed, $\psi((1, 0, 0, 0, 0)) = \varphi((1, 0, 0, 0, 0)) = 20$ and $\psi((0, 1, 0, 0, 0)) = \varphi((0, 1, 0, 0, 0)) = 2$.

(c) The model allows to "increase the size of the animals". For example, if we have a "very big dog" we can write

$$\psi(i(very\, big\, dog)) = \psi((0, 1, 0, 2, 0))$$

$$= \varphi((0, 1, 0, 0, 0)) + f((0, 0, 0, 2, 0)) = 2 + 2 \cdot f((0, 0, 0, 1, 0)) = 4.$$

(d) This model certainly allows the comparison of sizes of animals, even if they do not belong to the same species. For example, compare the size of a "big dog" and a "very small cat". This is given by

$$\left|\psi((0, 1, 0, 1, 0)) - \psi((0, 0, 1, 0, 2))\right|$$

$$= \left|\varphi((0, 1, 0, 0, 0)) + f((0, 0, 0, 1, 0)) - \varphi((0, 0, 1, 0, 0)) - 2\, f((0, 0, 0, 0, 1))\right|$$

$$= \left|2 + 1 - 1 - (2(-1))\right| = 4.$$

The examples presented above demonstrate two fundamental properties of our semantic embedding framework: its dynamic adaptability and context sensitivity. The framework's dynamism is evident in how it handles semantic compositions across different hierarchical levels. Beginning with basic categories (cow, dog, cat), the model incorporates modifiers (big, small, very) while maintaining meaningful numerical relationships. This adaptability allows the framework to capture subtle semantic variations without losing the underlying logical structure of size relationships.

The context sensitivity of the model is particularly noteworthy in three aspects. First, it preserves relative relationships within categories, as shown in the difference between "big cat" and "small cat". Second, it maintains coherent cross-category comparisons, enabling meaningful size evaluations between different animal species. Third, and perhaps most significantly, it handles compound modifications ("very big dog") in a way that respects both the base category's properties and the intensifying effect of modifiers.

Our model handles nonlinear feature interactions by combining linear and Lipschitz components. The linear part f handles direct comparisons in L-space for properties like size, while the nonlinear component ϕ uses a Lipschitz function with constant K to model complex object relationships. For example, ϕ assigns different size values to animals (cow = 20, dog = 2, cat = 1) in V-space, capturing nonlinear size relationships. We extend ϕ to other points using the McShane-Whitney formula while maintaining the Lipschitz property. The complete semantic projection $\psi = \phi \oplus f$ preserves these nonlinear interactions while ensuring mathematical tractability.

Thus, we have shown that this procedure allows to quantify the size of three different animals—using the antonym notions of "big" and "small"—without mixing linear and non-linear properties and without producing artifacts with no meaning in the mathematical model. Moreover, this construction can be used when the family of animals is extended, or even if other elements different than animals are introduced in the model (for example, "tree" or "rock"). This idea of introducing properties of nouns as "normal" elements of word embeddings is not, however, the only way to understand these models. We can consider them as properties, that is, as concepts whose meaning is evidenced by the

way they act on the set of nouns. This is not necessarily exclusive of the type of model we have already explained. We analyze this point of view in the next section.

4.3. Features on word embeddings and Lipschitz functionals

Note, however, that in the example model of the previous section, the adjectives "big" and "small" are introduced as terms included in the word embedding. The idea proposed in [10] is that adjectives can be represented as linear functionals on the space in which the word embedding is defined, rather than in the space itself. As we said, the use of such functionals is what is called semantic projections, and this notion is what we generalize here to Lipschitz functions as suggested in the previous section. In this section, we show how this can also be done using our mathematical structure.

The main idea is that, given a word embedding $i : W \hookrightarrow M$ in a metric space (M, d), we can increase its power as a semantic tool by adding new features, represented by Lipschitz real functionals $\varphi \in Lip(M)$. In this sense, we do not need to introduce the notions of "small" and "big" as elements in the word embedding but as elements of its dual space, that is, as a Lipschitz functional. Thus, in the example of the previous section, the notion of being "small" or "big" can be introduced as a Lipschitz function φ acting in the restricted word embedding $i : W \hookrightarrow (R^3, d)$, for $W = \{cow, dog, cat\}$ defined on the vectors $(1, 0, 0), (0, 1, 0)$, and $(0, 0, 1)$, respectively, and d is any metric (typically, the restriction of the Euclidean norm to this set of three vectors). Note that, following the idea of the previous section, the embedding can be done by composing again with the Lipschitz isometry $\iota : (i(W), d) \hookrightarrow (Æ(i(W)), \|\cdot\|_{Æ})$, to get the representation $\iota \circ i : W \hookrightarrow Æ(i(W))$. In this setting, φ is an element of the dual space $Lip(i(W))$ of $Æ(i(W))$. That is, it can be seen as a linear functional, but acting in the Banach space $Æ(i(W))$ instead of $\mathbb{R}^3$.

This Lipschitz functional $\varphi : \mathbb{R}^3 \to \mathbb{R}$ can then be freely defined as

$$\varphi((1, 0, 0)) = 6, \quad \varphi((0, 1, 0)) = 2, \quad \varphi((0, 0, 1)) = 1.$$

It can be noticed that, in this concrete case, the function φ could be considered as a linear functional acting in $\mathbb{R}^3$. However, this is not needed, and linearity refers to a more complex structure—the whole linear space $\mathbb{R}^3$—that does not make sense as an element of the model.

With this example, we wanted to show that the model can be fixed for representing just the three animals, and the property "size" can be introduced in a second step (as well as any other "adjective") as a Lipschitz function acting in a metric space with just three points.

4.4. Mathematical parameters and control formulas for the proposed mixed method

Once we have fixed features acting in the model as Lipschitz functionals, we have to fix the main bounds and error formulas for them. So, let us explain here the main equations that give the error bounds for the proposed model, based on Lipschitz-type inequalities and Euclidean residuals.

4.4.1. Adequacy to the metric structure of the model

To analyze the features of the concepts in the universe of the model, we need to define a characteristic parameter to measure how far the property follows the behavior of the distance. This means that, as the only structure we accept on the word embedding is the distance between its elements, a property—a Lipschitz function—is as adequate to the metric environment—the

distance—, as this parameter is close to a certain numerical value. Let us define this; we assume in what follows that the representation provided by the word embedding $i(W)$ is finite and $|W| = n$.. Assume that φ is not a constant index (otherwise, the results are trivial, so we give the values to the constants appearing below $Lip(\varphi) = 0$ and $Inv(\varphi) = Adq(\varphi) = \infty$.

The first assumption on a good Lipschitz index is Lipschitz inequality,

$$\left|\varphi(x_i) - \varphi(x_j)\right| \leq Lip(\varphi)\, d(x_i, x_j), \quad x_i, x_j \in i(M).$$

This gives a first Lipschitz parameter ($Lip(\varphi)$) providing information about the relation between the index and the metric. Notice that we can use it for all elements of $i(W)$ to get the inequality

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\varphi(x_i) - \varphi(x_j)\right|^2 \leq Lip(\varphi)^2 \sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2.$$

On the other hand, considering the converse inequality, we can define the inverse parameter $Inv(\varphi)$ by

$$Inv(\varphi)^2 = \Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\varphi(x_i) - \varphi(x_j)\right|^2\Big)^{-1} \cdot \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2\Big).$$

Finally, we define the adequacy parameter $Adq(\varphi)$ by

$$Adq(\varphi) = Lip(\varphi) \cdot Inv(\varphi).$$

This parameter $Adq(\varphi)$ gives a measure of how well the index φ fits the metric structure of the model. The following results shows how the extreme case when absolutely fits the metric, they coincide in the Lipschitz sense.

Proposition. Let $\varphi : i(W) \to \mathbb{R}$ be a non-constant Lipschitz index. Then, $Adq(\varphi) \geq 1$, and if $Adq(\varphi) = 1$, we have that

$$d(x_i, x_j) = Inv(\varphi) \cdot \left|\varphi(x_i) - \varphi(x_j)\right|, \quad x_i, x_j \in i(W).$$

Proof. Just using the definitions, we obtain the inequalities

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2 = Inv(\varphi)^2 \cdot \Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\varphi(x_i) - \varphi(x_j)\right|^2\Big)$$

$$\leq Inv(\varphi)^2 \cdot Lip(\varphi)^2 \cdot \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2\Big).$$

Thus, $Adq(\varphi) = Inv(\varphi) \cdot Lip(\varphi) \geq 1$. Now, let us assume that $Adq(\varphi) = 1$. Then

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2 = Inv(\varphi)^2 \cdot \Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\varphi(x_i) - \varphi(x_j)\right|^2\Big)$$

$$\leq Lip(\varphi)^2 \cdot Inv(\varphi)^2 \cdot \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2\Big) = \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2\Big).$$

We claim now that $d(x_i, x_j) \leq Inv(\varphi)\,|\varphi(x_i) - \varphi(x_j)|$ for every $i \neq j$. Otherwise, there are $i \neq j$ such that $Inv(\varphi)\,|\varphi(x_i) - \varphi(x_j)| < d(x_i, x_j)$ and so, there are two elements $x_k \neq x_l$ such that $Inv(\varphi)\,|\varphi(x_k) - \varphi(x_l)| > d(x_k, x_l)$. Thus,

$$d(x_k, x_l) < Inv(\varphi)\,|\varphi(x_i) - \varphi(x_j)| \leq Inv(\varphi) \cdot Lip(\varphi)\, d(x_k, x_l) = d(x_k.x_l),$$

a contradiction. This means that

$$d(x_i, x_j) \leq Inv(\varphi)\,|\varphi(x_i) - \varphi(x_j)| \quad for\ all \quad i \neq j,$$

which, together with

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2 = Inv(\varphi)^2 \cdot \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} \big|\varphi(x_i) - \varphi(x_j)\big|^2\Big)$$

gives $d(x_i, x_j) = Inv(\varphi)\,|\varphi(x_i) - \varphi(x_j)|$ for all $i \neq j$, and we get the result.

□

The proposition above establishes that when the adequacy index, calculated as described, is equal to one, the index φ can be directly determined by the metric, and vice versa. This suggests that the conceptual alignment between the index and the metric is ideal. In other words, comparing the index values of two semantic items yields their distance (up to a constant), indicating a perfect correlation between the semantic distance and the index comparison. Thus, the index fully captures the distance between items, and comparing these index values gives a direct estimate of the semantic distance, providing a measure of how well the index aligns with the metric structure of the space.

4.4.2. Non linearity deviation of the Lipschitz index

Suppose we have a word embedding into an Euclidean space of dimension n. Let $\widehat{\varphi}$ be the linear regression solution for the Lipschitz index φ. We are interested in measuring the error made when we represent the property with a linear functional rather than with a Lipschitz functional. Note that in this case $d = \|\cdot\|$, the Euclidean norm. The first thing we need to think about is that, given a *finite* model for n semantic items $\{s_1, ..., s_n\}$, each property for which we have a numerical estimate $i \mapsto e_i$ can be written directly with a Lipschitz functional. That is, the assignment $s_i \mapsto \varphi(s_i) = e_i$ always defines a Lipschitz function. Of course, the Lipschitz constant can get a very large value, and the adequacy index defined in the previous section $Adq(\varphi)$ could be very bad ($Adq(\varphi) >> 1$). In this section we want to present the mathematical tool to measure the linearization error, that is, *how far is φ of being linear*.

In order to do this, we can measure two different quantities.

(1) The regression error is the first related error. As we said, given a set of numerical estimates of a certain feature, we can always find a Lipschitz function φ such that $\varphi(s_i) = e_i$. Let $\widehat{\varphi}$ be the result of the linear regression of the data $\{(s_i, e_i) : i = 1, ..., n\}$, that is taken to be the best linear approximation to φ. The quadratic error can be defined in terms of the residues, that is

$$\varepsilon_1(\varphi)^2 = \sum_{i=1}^{n} \big|\widehat{\varphi}(x_i) - \varphi(x_i)\big|^2 = \sum_{i=1}^{n} \big|\widehat{\varphi}(x_i) - e_i\big|^2.$$

Therefore, it coincides with the usual quadratic error of the corresponding linear regression. Of course, $\varepsilon_1 = 0$ if and only if φ is linear.

(2) The second-order error is given by the difference of the residues considered as Lipschitz functions; it provides the quadratic error adapted to a Lipschitz-type inequality, as we show in the next proposition.

$$\varepsilon_2(\varphi) = \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| \widehat{\varphi}(x_i) - \varphi(x_i) - (\widehat{\varphi}(x_j) - \varphi(x_j)) \right|^2 \Big)^{1/2}.$$

Due to the fact that $\widehat{\varphi}$ is a Lipschitz function, we have the inequalities

$$Lip(\varphi) \leq Lip(\varphi - \widehat{\varphi}) + \|\widehat{\varphi}\| \quad \text{and} \quad \|\widehat{\varphi}\| \leq Lip(\varphi - \widehat{\varphi}) + Lip(\varphi).$$

Let us see that if $\varepsilon_2 = 0$, we have that φ is an affine function. Note that, by assumption, $\widehat{\varphi}$ can be written as a linear map plus a fixed point b, that is, there is a vector $v \in \mathbb{R}^n$ such that $\widehat{\varphi}(x_i) = v \cdot x_i + b$. Indeed, in this case we have that for all $i = 1, ..., n$,

$$\left| \widehat{\varphi}(x_i) - \varphi(x_i) - (\widehat{\varphi}(x_j) - \varphi(x_j)) \right| = 0,$$

We get $\varphi(x_i) = \widehat{\varphi}(x_i) + (\varphi(x_1) - \widehat{\varphi}(x_1)) = \widehat{\varphi}(x_i) + a$, where $a = \varphi(x_1) - \widehat{\varphi}(x_1)$ is a fixed point. Thus, $\varphi(x_i) = \widehat{\varphi}(x_i) + a = v \cdot x_i + (b + a)$, an affine function.

The following result shows that the second order error ε_2 has a natural bound in terms of the 2−norm of the metric matrix, and the addition of the corresponding norm and Lipschitz constants. This error provides an idea of the sizes of the deviation of the Lipschitz constant of φ and its linear approximation $\widehat{\varphi}$. It can be considered as a Lipschitz map $M : \mathcal{M}_{n\times n} \to \mathbb{R}$; its Lipschitz-type inequality, when the matrix considered is the metric matrix, is given by the following result. Let us write $\|D\|_2$ for this norm, that is, $\|D\|_2 = \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - x_j\|^2 \Big)^{1/2}$.

Proposition. If D is the Euclidean metric matrix, the inequality

$$\varepsilon_2(\varphi) \leq (\|\widehat{\varphi}\| + Lip(\varphi)) \, \|D\|_2$$

holds for every Lipschitz map φ and every Lipschitz approximation $\widehat{\varphi}$ to it.

Proof. The following straightforward computations give the result.

$$\begin{aligned}
\varepsilon_2 &= \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| \widehat{\varphi}(x_i) - \varphi(x_i) - (\widehat{\varphi}(x_j) - \varphi(x_j)) \right|^2 \Big)^{1/2} \\
&= \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| (\widehat{\varphi}(x_i) - \widehat{\varphi}(x_j)) + (\varphi(x_j) - \varphi(x_i)) \right|^2 \Big)^{1/2} \\
&\leq \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| \widehat{\varphi}(x_i) - \widehat{\varphi}(x_j) \right|^2 \Big)^{1/2} + \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| \varphi(x_j) - \varphi(x_i) \right|^2 \Big)^{1/2} \\
&\leq \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \|\widehat{\varphi}\|^2 \, \|x_i - x_j\|^2 \Big)^{1/2} + \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} Lip(\widehat{\varphi})^2 \, \|x_i - x_j\|^2 \Big)^{1/2}.
\end{aligned}$$

□

Essentially, a small error ϵ_2 indicates that the Lipschitz functional representing the concept follows an approximately linear trend, meaning there is an (almost) linear ordering of the elements according to the property represented by the functional. The proposition quantifies also this relation in terms of the linear and Lipschitz norm, as well as the norm of the distance matrix.

5. Nonlinear word embeddings and semantic projections: Nouns+adjectives model

In the following sections, we focus on a simplified scenario of the general model previously described. We will consider semantic projections as a means of representing properties—typically represented as adjectives—that affect a given universe of nouns, thereby indicating concepts. These will form the elements of the metric space underlying the model. Additionally, we will introduce properties concerning the set of nouns (adjectives) as Lipschitz functions. It is important to note that while the elements of the metric model are often represented as vectors in examples, the linear structure is not utilized when considering Lipschitz functions for adjective representation instead of linear functionals. In fact, the residual error measures how far the property represented by a Lipschitz function deviates from linearity.

For our example, we have used tools available in the DeflyCompass platform in its trial version, which is designed to calculate semantic projections using various methods. It is worth recalling that any semantic projection is defined using different criteria, but all yield a real number $P_U(t) \in [0, 1]$ when a universe U and a term t are fixed. We compute the semantic projections associated with each element of the universe separately, calculating $P_w(t)$ for every $w \in U$. Broadly speaking, $P_w(t)$ represents the extent to which the meaning of the term t is related to the meaning of w within the context established by U and the specific definition of the semantic projection PP.

5.1. A concrete example: The term "precious" in a universe of metals

We are operating within the context of the relationships of nouns and adjectives to compare with the typical understanding of semantic projections. To facilitate comprehension of our ideas, we present a very simple case. We consider a universe U composed of five words representing precious materials, alongside other metals that are less valuable. Concretely,

$$U = \{\text{“}diamond\text{”}, \text{“}gold\text{”}, \text{“}silver\text{”}, \text{“}copper\text{”}, \text{“}iron\text{”}\}.$$

We consider two examples of properties that can be applied to all elements of U. The first one, the noun "price" and the second, the adjective "precious". Note that, although "price" is a noun, the semantic role of "price" here in comparison with the U elements is a feature, in the sense that it is connected to the idea of how valuable the object is.

To analyze the relationships, we propose two different semantic projections, the first based on Google Search and the second using the cosine similarity defined in the Word2Vec word embedding.

1) Semantic projection based on Google Search: the semantic projection P^G is defined as

$$P^G_w(t) = \frac{\text{number of documents in which } t \text{ and } w \text{ appear together in Google Search}}{\text{number of documents in which } t \text{ appears in Google Search}}.$$

The results can be seen in the next Figures 1 and 2. As a first example, the first one represents the values for the term "price" (Figure 1), the second one shows the result for the adjective "precious" (Figure 2).

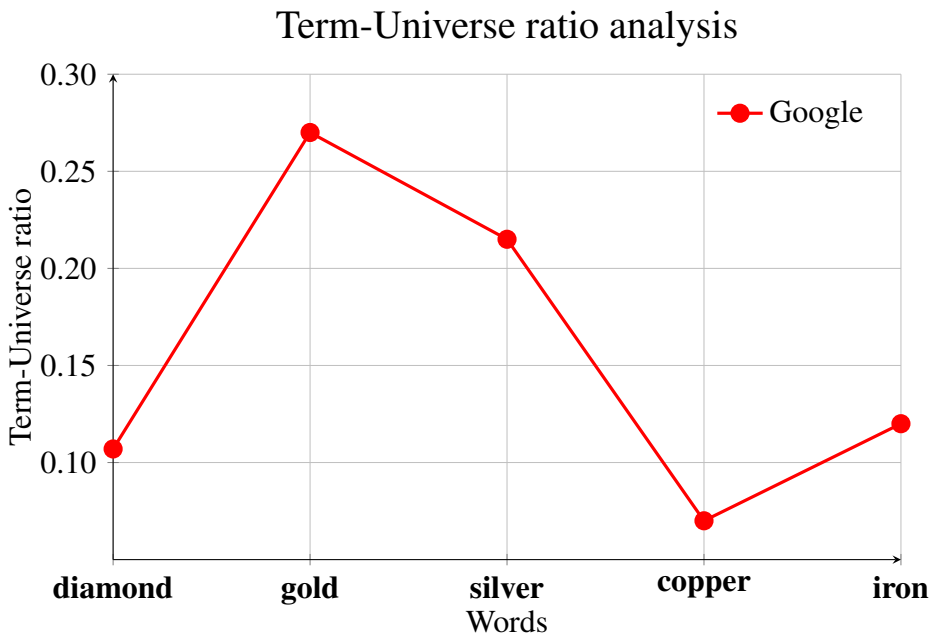

Figure 1. Google semantic projection of the noun "price" on the universe of metals+diamond.

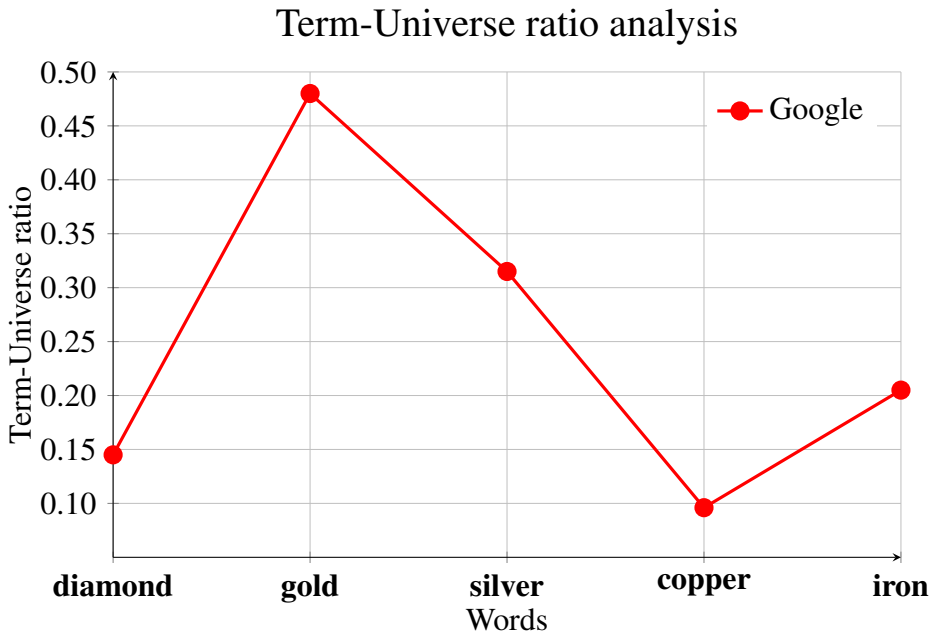

Figure 2. Google semantic projection of the adjective "precious" on the universe of metals+diamond.

Regarding the term "precious", which will be further analyzed in the next section, Figure 3 represents a part of the universe U around this term provided by the embedding projector that can be found in http://projector.tensorflow.org/.

Figure 3. Representation centered on the term "precious" of the nearest words using the Word2Vec word embedding, which includes a relevant part of the universe U.

2) Semantic projections defined using Word2Vec:
We can define a new semantic projection using the cosine similarity in the word embedding provided by Word2Vec of the words in U and the term "precious". Let us write $CS(a, b)$ for the cosine similarity between the vectors representing the words a and b, that measures a semantic distance between these words in the word embedding. Then, the value of a semantic projection $P_w^{cos}(t) \in [0, 1]$ for $w \in U$ and the term t can be defined as

$$P_w^{cos}(t) = \frac{1}{2}(1 + CS(w, t)).$$

This formula transforms cosine similarity (CS), which ranges from -1 to 1, into a $[0, 1]$ range, making it suitable as a semantic projection. The behavior of this transformation can be understood through its key properties: In the case of opposite vectors where $\text{CS}(w, t) = -1$, we obtain $P_w^{\cos}(t) = 0$, indicating maximum semantic dissimilarity. For orthogonal vectors with $\text{CS}(w, t) = 0$, the projection yields $P_w^{\cos}(t) = 0.5$, representing a neutral semantic relationship. Finally, when the vectors are identical and $\text{CS}(w, t) = 1$, we get $P_w^{\cos}(t) = 1$, denoting maximum semantic similarity.

As the reader can see, the way these properties are computed, in both semantic projections, does not suggest the use of linear relations to model the adjective "precious".

5.2. *Lipschitz vs linear model in the example*

Let us now show all the machinery we have introduced in the previous section to analyze the case of the adjective "precious" and the reduced universe that comes from the one above but removing the term "diamond", that is, in the case

$$U = \{gold, silver, copper, iron\}.$$

We can use the computing platform for Word2Vec (Figure 3) to find the Euclidean distances between the terms of the universe. The resulting metric matrix is

$$D = \begin{bmatrix} 0 & 0.308 & 0.545 & 0.640 \\ 0.308 & 0 & 0.452 & 0.501 \\ 0.545 & 0.452 & 0 & 0.447 \\ 0.640 & 0.501 & 0.447 & 0 \end{bmatrix},$$

where the order in the matrix is the one given in U.

For the aim of clarity, we have computed a projection of this representation in a 3-dimensional space by means of the coordinates

$$\begin{gathered} \text{Gold: } (0.863, 0.579, 0.682), \\ \text{Silver: } (0.657, 0.806, 0.70), \\ \text{Copper: } (0.409, 0.617, 0.382), \\ \text{Iron: } (0.232, 0.553, 0.788), \end{gathered}$$

which preserve the Euclidean distance among them. This can be seen in Figure 4.

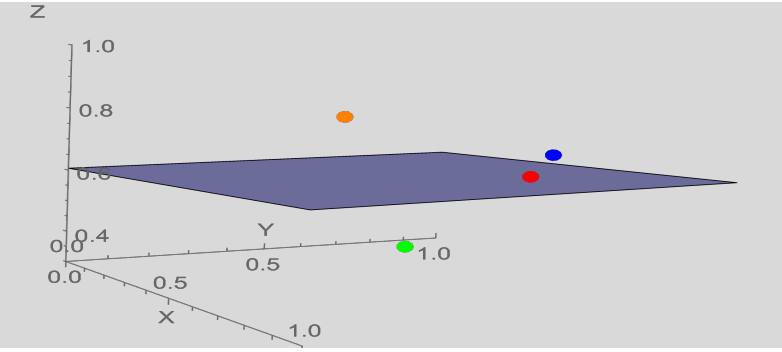

Figure 4. Representation of the four points of the reduced universe.

Let us argue now how the adjective "precious" operates on the reduced universe U. In order to make the representation easily understandable, we take only the first coordinate of each metal, following the order in U, and consider the semantic projection provided by P^G and computed above. This gives the following set of (x, y) coordinates

$$\{(0.863, 0.27), (0.657, 0.22), (0.409, 0.07), (0.232, 0.12)\}.$$

In order to represent the function in the direction of increasing value, the order in the representation (Figure 5), starting from the lowest values of x, is: iron, copper, silver, and gold.

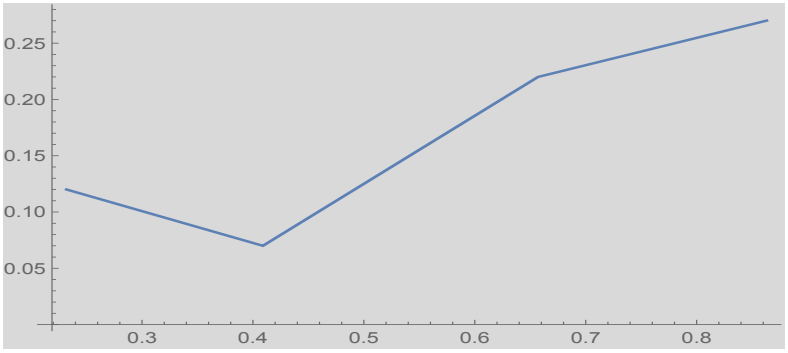

Figure 5. Representation of the semantic projection of the adjective "precious" as a function of the first coordinate of the four points of the reduced universe.

Figure 5 clearly shows that there is no linear dependence on the vector representation of the metals with the value of the semantic projection that represent the adjective "precious". The best linear fit can be seen in Figure 6. The formula of the line is $y = 0.0135 + 0.2897\,x$.

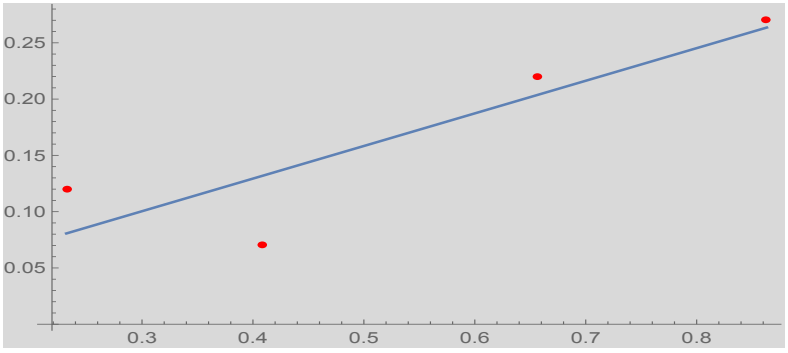

Figure 6. Linear fit of the points representing the metals defined by the term "precious".

We can use the results on the adequacy of the linear assumption shown in Section 4.4. On the one hand, we can compute the adequacy index. In order to do this, we first calculate the Lipschitz constant of the function φ, that gives $Lip(\varphi) = 0.367$. Together with the calculus of the inverse of the relative error sums $Inv(\varphi)$, that gives $Inv(\varphi)^2 = (1/0.2) \cdot 2.913 = 14.565$, we find the value $Adq(\varphi) = Lip(\varphi) \cdot Inv(\varphi) = 0.367 \cdot 14.565 = 5.345$, meaningfully bigger to 1. Recall that it equals 1 in case the map is linear.

The error ε_2 regarding the residues of the differences between the linear approximation and the true Lipschitz functions can be directly computed, and gives the value $\varepsilon_2 = \sqrt{0.0456} = 0.2135$. Although this number is only meaningful when compared to other values, a relatively small value can be interpreted as a good fit of the Lipschitz function to its linear approximation. In other words, while the Lipschitz function provides a more accurate fit to the points (as it coincides with all sample points), it remains relatively similar to a linear functional, which is a semantic projection as understood in the conventional linear-based model.

5.3. Discussion: Scalability and sentence comparison in large-scale semantic analysis

We have shown in simple examples the ability of the method to enhance both the meaning of the elements within the model and the way these elements are interpreted. However, these tools are designed for use in more complex semantic environments, where the objects to be compared involve a higher level of complexity, such as large sets of words or complete sentences.

In our approach, scalability is primarily a technical rather than a conceptual issue. When dealing with a large model containing thousands of semantic items, Lipschitz functionals can be defined for all of them using the same method. This is because there are no restrictions on assigning appropriate values to these functionals when the set is finite, as is the case with models based on linear functionals. Any function defined on a finite set of a metric space can always be perfectly adapted to a Lipschitz functional, although the Lipschitz constant could become quite large, which, as explained in previous sections, may lead to a loss of accuracy in the results, so this has to be controlled. If the set of words includes items of different types, some of which are of a completely unrelated nature, an arbitrary constant value (such as 0) can be assigned to those words for which the property in question has no meaning (for instance, the adjective "color" in a semantic context involving ethical concepts). This allows for the Lipschitz constant to remain controlled, as the function's value for semantic items where the concept is irrelevant would not affect the outcome. This flexibility is another advantage over linear modeling, where such an adjustment cannot be easily applied.

It is also important to note that different criteria are modeled by independent Lipschitz functions, which does not limit the representation's nature. Two sentences composed of elements containing diverse information can be compared without altering each other's meaning, as these criteria are handled by separate Lipschitz functions. For complex linguistic structures, a general procedure for defining Lipschitz functions that allow for meaningful comparisons can be devised according to the nature of the property being modeled. For example, a composite index, formed by adding the pairwise comparisons of semantic items within each complex structure (such as an entire sentence), could serve this purpose. Metric comparisons for obtaining the Lipschitz inequality between such complex elements can again be performed by computing the distances between all the items in each semantic structure.

These considerations require further analysis, and more detailed constructions will be necessary as the complexity of semantic environments increases. These questions remain open for future research in this area.

6. Conclusions

Although linearity is a difficult tool to use for semantic embeddings $W \hookrightarrow E$, it is sometimes convenient for certain terms in semantic environments. The notion of semantic projection has proven to be a useful tool for quantifying properties in the model, but in its original definition (see [10]), it has to be represented as a linear real-valued map of E^*. However, the linear structure is sometimes confusing as a model feature, and some of the consequences that follow from its use have to be omitted. This can be overcome if the requirement for real functions playing the role of semantic projections is only to be Lipschitz, and not necessarily linear. This implies some compatibility with the metric—the fundamental underpinning of word embeddings—but does not make dependence on the linear structure of E mandatory.

This insight leads us to propose a new word embedding as a mapping from the original semantic environment to a direct sum of a subspace L of E and a free space $AE(V)$, where $E = L \oplus V$. A semantic projection is then modeled as a functional of the space $L^* \oplus Lip(V)$, that is, a direct sum $f \oplus \varphi$ of a linear map f and a Lipschitz map φ.

These ideas culminate in a systematic framework for constructing nonlinear semantic embeddings, where semantic projections of the form $f \oplus \varphi$ serve to quantify the properties of model terms. The development of this procedure, including its governing mathematical parameters and demonstration through a concrete applied case, constitutes the principal contribution of this work. The theoretical foundations established here extend beyond our current focus on text analysis and word embeddings, suggesting a broader mathematical framework with far-reaching implications. The semantic projection architecture, particularly the interplay between linear and nonlinear components through $\psi = \phi \oplus f$, provides a natural pathway to diverse data modalities. This versatility invites applications in image analysis, audio processing, and other structured or unstructured data domains, while preserving the mathematical rigor essential for complex feature interactions. As the field of multimodal analysis and representation learning continues to evolve, our framework offers a robust foundation for developing unified semantic models.

Author contributions

Pedro Fernández de Córdoba: Validation, Conceptualization, Investigation, Writing-review & editing, Supervision; Carlos A. Reyes Pérez: Software, Formal analysis, Writing-original draft, Visualization; Enrique A. Sánchez Pérez: Formal analysis, Writing-original draft, Conceptualization, Investigation. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

We would like to acknowledge funding from the Generalitat Valenciana (Spain) through the PROMETEO CIPROM/2023/32 grant.

Conflict of interest

The authors declare that there are no conflict of interest.

References

1. C. D. Aliprantis, K. C. Border, *Infinite Dimensional Analysis*, 3 Eds., Germany: Springer, 2006.
2. R. F. Arens, J. Eels Jr., On embedding uniform and topological spaces, *Pacific J. Math.*, **6** (1956), 397–403. https://doi.org/10.2140/pjm.1956.6.397

3. R. Arnau, J. M. Calabuig, E. A. Sánchez Pérez, Representation of Lipschitz Maps and Metric Coordinate Systems, *Mathematics*, **10** (2022), 3867. https://doi.org/10.3390/math10203867
4. M. Baroni, R. Zamparelli, Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, 1183–1193.
5. G. Boleda, Distributional semantics and linguistic theory, *Ann. Rev. Linguist.*, **6** (2020), 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303
6. S. Clark, Vector space models of lexical meaning, In: *The Handbook of Contemporary Semantics*, Malden: Blackwell, 2015, 493–522.
7. Ş. Cobzaş, R. Miculescu, A. Nicolae, *Lipschitz functions*, Berlin: Springer, 2019.
8. J. Dai, Y. Zhang, H. Lu, H. Wang, Cross-view semantic projection learning for person re-identification, *Pattern Recognit.*, **75** (2018), 63–76. http://dx.doi.org/10.1016/j.patcog.2017.04.022
9. K. Erk, Vector space models of word meaning and phrase meaning: A survey, *Lang. Linguist. Compass*, **6** (2012), 635–653. http://dx.doi.org/10.1002/lnco.362
10. G. Grand, I. A. Blank, F. Pereira, E. Fedorenko, Semantic projection recovers rich human knowledge of multiple object features from word embeddings, *Nat. Hum. Behav.*, **6** (2022), 975–987. https://doi.org/10.1038/s41562-022-01316-8
11. N. J. Kalton, Spaces of Lipschitz and Hölder functions and their applications, *Collect. Math.*, **55** (2004), 171–217.
12. J. L. Kelley, General Topology, *Graduate Texts in Mathematics*, New York: Springer, 1975.
13. A. Lenci, Distributional models of word meaning, *Ann. Rev. Linguist.*, **4** (2018), 151–171. http://dx.doi.org/10.1146/annurev-linguistics-030514-125254
14. O. Levy, Y. Goldberg, Neural word embedding as implicit matrix factorization, *Adv. Neural Inf. Proc. Syst.*, 2014, 2177–2185.
15. H. Lu, Y. N. Wu, K. J. Holyoak, Emergence of analogy from relation learning, *Proc. Natl. Acad. Sci. U. S. A.*, **116** (2019), 4176–4181. http://dx.doi.org/10.1073/pnas.1814779116
16. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, *Adv. Neural Inf. Proc. Syst.*, 2013, 3111–3119.
17. J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, 2014. Available from: `https://nlp.stanford.edu/projects/glove/`.
18. N. Weaver, *Lipschitz Algebras*, Singapore: World Scientific Publishing Co., 1999.
19. Y. Xian, S. Choudhury, Y. He, B. Schiele, Z. Akata, Semantic projection network for zero-and few-label semantic segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 8256–8265. http://dx.doi.org/10.1109/CVPR.2019.00845
20. L. A. Zadeh, A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges, *J. Cybern.*, **2** (1972), 34–34.