Information Theory–based Compositional Distributional Semantics

Enrique Amigó
Universidad Nacional de Educación a Distancia (UNED)
enrique@lsi.uned.es

Alejandro Ariza-Casabona
CLiC - UBICS, Universitat de Barcelona
alejandro.ariza14@ub.edu

Víctor Fresno
Universidad Nacional de Educación a Distancia (UNED)
vfresno@lsi.uned.es

M. Antònia Martí
CLiC - UBICS, Universitat de Barcelona
amarti@ub.edu

In the context of text representation, Compositional Distributional Semantics models aim to fuse the Distributional Hypothesis and the Principle of Compositionality. Text embedding is based on co-ocurrence distributions and the representations are in turn combined by compositional functions taking into account the text structure. However, the theoretical basis of compositional functions is still an open issue. In this article we define and study the notion of Information Theory–based Compositional Distributional Semantics (ICDS): (i) We first establish formal properties for embedding, composition, and similarity functions based on Shannon's Information Theory; (ii) we analyze the existing approaches under this prism, checking whether or not they comply with the established desirable properties; (iii) we propose two parameterizable composition and similarity functions that generalize traditional approaches while fulfilling the formal properties; and finally (iv) we perform an empirical study on several textual similarity datasets that include sentences with a high and low lexical overlap, and on the similarity between words and their description. Our theoretical analysis and empirical results show that fulfilling formal properties affects positively the accuracy of text representation models in terms of correspondence (isometry) between the embedding and meaning spaces.

Action Editor: Kevin Duh. Submission received: 29 September 2021; revised version received: 15 July 2022; accepted for publication: 21 July 2022.

https://doi.org/10.1162/coli_a_00454

1. Introduction

The representation of text meaning is a fundamental issue in Natural Language Processing (NLP). It involves encoding natural language in a way that can be handled by information management systems. It is one of the main bottlenecks in Textual Information Access, Text Mining, Dialogue Systems, and so forth. The conflict between compositionality and contextuality (Frege's principles) exhibits a tension between the meaning representation paradigms, namely, symbolic and distributional (Maruyama 2019).

The Principle of Compositionality states that the meaning of a whole is a function of the meaning of its parts and the syntactic way in which they are combined together. This principle is the main foundation of the symbolic representation paradigm, which attempts to relate language to propositional logic via extended semantic references (word referential meaning) and grammars that capture the language structures. For instance, in the symbolic paradigm, *"table"* and *"sitting"* are associated with their references (icons in Figure 1a) and the utterance *"sitting at a table"* is the result of being combined by means of a syntactic structure.

On the other hand, The Principle of Contextuality states that the meaning of words and utterances[1] is determined by their context. This principle supports the distributional representation paradigm, in which the meaning of linguistic items is inferred from their context of use. Rather than symbols, words or utterances are represented as points in a continuous vector space (embeddings; i.e., vectors in Figure 1b). For instance, the words *"table"* and *"sitting,"* and the utterance *"sitting at a table"* are projected in the representation space according to the textual context in which they usually appear. In contrast to the symbolic paradigm, the distributional paradigm interprets the meaning space as a continuous, that is, there exists a graded scale of meaning representations between *"sitting," "table,"* and *"sitting at a table."*

Over the last decade, distributional NLP approaches have undoubtedly been the predominant basis of NLP applications thanks to their predictive power over a sufficient amount of textual data. In particular, Neural Language Models, such as Generative Pre-trained Transformer (GPT) (Radford et al. 2019) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019), pre-trained over a huge amount of parameters and text collections, have demonstrated high performance on specific tasks with limited supervised data. However, in terms of meaning representation, there is considerable consensus in the literature that one of the main limitations of the distributional paradigm is its lack of **systematicity** (Johnson 2004; Talmor et al. 2020; Pimentel et al. 2020; Goodwin, Sinha, and O'Donnell 2020; Hupkes et al. 2020; Bender and Koller 2020). Robert Cumming defined language systematicity as follows (Cummins 1996, page 591): *A system is said to exhibit systematicity if, whenever it can process a sentence, it can process systematic variants, where systematic variation is understood in terms of permuting constituents or (more strongly) substituting constituents of the same grammatical category*. Briefly, while in the symbolic paradigm the productive rules defining grammars admit infinite combinations and permutations, the distributional paradigm is somewhat conditioned by the word sequences observed in the training corpus.

1 In this article we will use the terms *word* and *utterance*. The notion of *word* used in this article can be generalized to any atomic element on which composition or distributional analysis is applied by the computational linguistic model. This ranges from morphemes in synthetic languages such as German to multiwords or collocations. On the other hand, we will use the term *utterance* to refer to the compound expressions we want to represent.

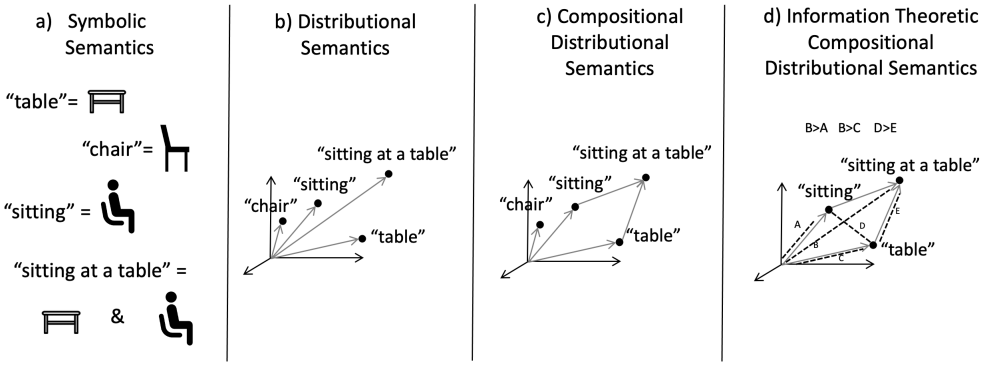

Figure 1
Representation paradigms in computational linguistics.

Because distributional and compositional representation approaches present complementary properties (contextuality vs. systematicity), Maruyama (2019) raised the need for a *Kantian Synthesis*. The hypothesis is that the meaning of linguistic forms is given by a reflow between their context and components. In this line, a number of authors have proposed adding a compositional layer over distributional representations. These models are referred to as Compositional Distributional Semantic models (Mitchell and Lapata 2010; Arora, Liang, and Ma 2017; Clark and Pulman 2007; Coecke, Sadrzadeh, and Clark 2010). As shown in Figure 1c, in the distributional compositional paradigm, the utterance *"sitting at a table"* is not inferred from the context in which it appears, but by a composition function applied on the distributional representations of *"sitting"* and *"table."* In other words, context of words determines their meaning. At the same time, the meaning of the utterance is given by the words that compose it and compositional operators. This allows us to exploit the compositionality systematicity on utterances whose length does not allow us to infer usage statistics.

The main goal of this article is the formal study of embedding, composition, and similarity functions in the context of Compositional Distributional representation models, allowing for the analysis of the problem independently of the task in which the representation is applied. For this purpose, we formalize the notion of *Information Theory–based Compositional Distributional Semantics*, the contribution of which to Distributional Compositional models consists of establishing formal constraints for embedding, composition, and similarity functions (Figure 1d). These constraints are based on Shannon's concept of Information Content (IC), establishing a link between distributional semantics, compositional language structures, and Information Theory.

We apply the following methodology. We first establish formal properties for embedding, composition, and similarity functions based on Shannon's Information Theory. Then we analyze the existing approaches under this prism, checking whether or not they comply with the established desirable properties. After this, we propose two parameterizable composition and similarity functions that generalize traditional approaches while fulfilling the formal properties. Finally, we perform an empirical study on several textual similarity datasets that include sentences with a high and low lexical overlap, and on the similarity between words and their description. Our theoretical analysis and empirical results show that fulfilling formal properties affects positively

the accuracy of text representation models in terms of correspondence between the embedding and meaning spaces.

The article is structured as follows. Section 2 analyzes the main existing semantic approaches from a historical perspective and reviews Compositional Distributional Semantics models in more detail. Sections 3 and 4 describe the proposed theoretical framework for Information Theory–based Compositional Distributional Semantics (ICDS). Section 5 describes the proposed generalized similarity and composition functions, their properties, and their connection with existing functions. Section 6 displays our experiments and results. Finally, we draw our conclusions in Section 7.

2. Related Work

2.1 Text Representation Paradigms

To get an overview of the different text representation paradigms, we will consider the following set of properties.

SYSTEMATICITY: Whenever the representation is able process a sentence, it can process systematic variants, where systematic variation is understood in terms of permuting constituents or (more strongly) substituting constituents of the same grammatical category.

USAGE CONTEXT: The meaning representation is sensitive to the contexts in which the expression appears. There is a consensus among psycholinguists that the semantics of the basic units of language (words, collocations) is determined by their use (Wittgenstein 1953). Put in computational linguistic terms, this is known as the Distributional Hypothesis (Harris 1954; Firth 1957).

CONTINUITY: There is a multidimensional continuous space where the meaning representations are mapped. The distributional approach offers us a way to capture the full continuum of meaning (Erk 2009; Gladkova and Drozd 2016). Moreover, this space should be **isometric** with respect to the meaning affinity of expressions, that is, it should reflect semantic similarities within the space of embeddings. The continuity property offers important advantages, such as the ability to generalize by proximity in space and to incorporate dimensionality reduction mechanisms, and so on (Landauer 1997).

INFORMATION MEASURABILITY: The representation system includes a function that measures the amount of information contained in the represented utterance. In general, the amount of information is measured in terms of specificity or likelihood, in concordance with Shannon's notion of Information Content ($I(x) = -\log(P(x))$). This links the representation system with the notion of language model, and is a key aspect in representation models (Zhai 2008) and textual similarity measures (Lin 1998).

In this section, we will show that, since the beginning of NLP, these properties have been captured or lost as new paradigms have been established. The loss of properties in each paradigm was compensated for by the use of different techniques. Table 1 summarizes this historic evolution.

The first semantic models were based on logicist approaches assuming the Principle of Compositionality (e.g., via Preference Semantics [Wilks 1968]). These models establish a univocal relationship between a symbol (or a set of symbols) in context and its meaning (Boleda and Erk 2015). The space of senses is given by ontological

Table 1
Semantic representation approaches and their properties. The table includes techniques used to mitigate drawbacks.

		SYSTEMATICITY	USAGE CONTEXT	CONTINUITY	INFORMATION MEASURABILITY
Symbolic Paradigm	Complying:	✔	✗	✗	✗
	Approach:		WSD Prob. grammars	Ontological similarity	Ontological deepness
Vector Space Model	Complying:	✗	✗	✔	✗
	Approach:	Tensor-based approaches	Dimensionality reduction		Feature weighting
Count-based Language Models	Complying:	✗	✔	✗	✔
	Approach:	Independence assumption		Similarity between probabilities	
Neural Language Models	Complying:	✗	✔	✔	✔
	Approach:	Vector operators			
Compositional Distributional Approaches	Complying:	✔	✔	✔	✔

resources such as WordNet (Fellbaum 1998). The symbolic paradigm is inherently systematic, since it connects with propositional logic. The other aspects are only partially captured. First, the compositional paradigm is fundamentally based on rules, not on the distributional features of language (USAGE CONTEXT); however, the context of use is partially captured by techniques such as Lexical Sense Disambiguation (Navigli 2009) that maps words to senses according to their context, and Probabilistic Grammars (Sekine and Grishman 1995) trained on text corpora. Although the semantic space in the symbolic paradigm is discrete, the lack of CONTINUITY is partially mitigated by means of similarity metrics between nodes within an ontology, for example, via Conceptual Density (Agirre and Rigau 1996) in WordNet. The lack of INFORMATION MEASURABILITY at the lexical level within these models has also been addressed via the deepness of nodes (specificity) in the hierarchical ontology (Seco, Veale, and Hayes 2004).

From the 1980s onward, the Statistical Paradigm, whose representation system was fundamentally based on the Vector Space Model (VSM) (Salton and Lesk 1965), gained strength. Under this paradigm, texts are represented as a bag of independent words, disregarding word order and grammar. Consequently, SYSTEMATICITY is not captured, since the structures of the language are ignored. Some approaches incorporate a certain level of SYSTEMATICITY by means of objects more complex than simple vectors, such as tensors (Baroni and Lenci 2010; Padó and Lapata 2007; Turney 2007). The natural language property that the VSM incorporates is CONTINUITY on which similarity measures are applied, such as cosine, dot product, or Euclidean distance. Because the original VSM assumes word independence, it does not capture USAGE CONTEXT. This lack is tackled via dimensionality reduction approaches based on the Distributional Hypothesis, such as Latent Semantic Indexing (Deerwester et al. 1990)

or Latent Dirichlet Allocation (Blei et al. 2003).[2] Regarding INFORMATION MEASURABILITY, the Term Frequency–Inverse Document Frequency (TF-IDF) term weighting function and its variants (such as projection functions that are within vector space) approaches the probability of the represented utterance and have a direct connection with the notion of Information Content in Shannon's theory (Robertson 2004).

The next generation of meaning representation consisted of the Count-based Language Models. In these models, texts are represented as word sequences and their probability distribution (Andreas, Vlachos, and Clark 2013). INFORMATION MEASURABILITY is captured via **perplexity**, a concept directly based on Shannon's notion of Information Content. Language Models are mainly usage-oriented, since perplexity is based on the frequency of words in given contexts. Texts are represented as a whole (holistic) in terms of the word sequence and its probability. Consequently, the lack of SYSTEMATICITY is a limitation of Language Models because the probability of a word sequence cannot be inferred from the probability of its parts (Bengio et al. 2003). Just like in the VSM, a way of mitigating this lack consists in assuming statistical independence and applying the product of probabilities; this is the case of the *n*-gram model (Brown et al. 1993). Regarding semantic CONTINUITY, Language Models represent texts as token sequences rather than a point in a continuous space. In addition, they do not provide a direct notion of similarity between word sequences, but between different Language Models (e.g., Kullback-Leibler divergence) or sequences vs. Language Models (perplexity).

Lastly, the most successful approaches over the last few years are supported by Neural Language Models, which are based on neural networks pre-trained over a huge text corpus on the basis of USAGE CONTEXT (Devlin et al. 2019; Brown et al. 2020). Neural Language Models combine the properties of VSMs and Language Models. On the one hand, text embedding is given by the activation state of neural network inner layers. Therefore, as in the VSM, the representation space is a continuum, allowing generalization. Just like traditional Language Models, the network is trained to predict words, estimating a probabilistic distribution of word sequences given their context. In consequence, Neural Language Models capture both USAGE CONTEXT and INFORMATION MEASURABILITY. Their limitation is still SYSTEMATICITY. Finally, the Compositional Distributional paradigm covers all properties and is discussed in detail in the next section.

In Table 1, a summary of the semantic models and the linguistic features of language that they address is shown. In view of this table, we can conclude that CONTINUITY, INFORMATION MEASURABILITY, USAGE CONTEXT, and SYSTEMATICITY are complementary desirable properties that have been addressed from different paradigms, although covering all of them simultaneously remains a challenge for the Natural Language Processing community.

2.2 Distributional Semantic Representation Models: From Word Embeddings to Transformers

In 2008, Collobert and Weston (2008) demonstrated that word embeddings generated from a suitably large dataset carry syntactic and semantic meaning and improve the performance on subsequent NLP tasks. However, the hype regarding neural encoders

2 Note that dimensionality reduction compresses the representation by taking into account the interdependence between terms.

really started in 2013 with the distributional semantic model known as Skip-gram with Negative Sampling (SGNS, Word2Vec package) (Mikolov et al. 2013). Some authors proposed the extension of this approach to represent longer linguistic units, such as sentences (Kiros et al. 2015) or documents (Le and Mikolov 2014; Kenter, Borisov, and de Rijke 2016). Static Neural Models (also referred to as Non-Contextual Neural Models) such as SGNS or GloVe (Pennington, Socher, and Manning 2014) optimize the correspondence between the scalar product of embeddings and their distributional similarity (Mutual Information) (Levy and Goldberg 2014; Le and Mikolov 2014; Arora et al. 2016). Assuming the distributional hypothesis (similar words appear in similar contexts), this ensures a certain isometry between the embedding space and meanings.

These embedding approaches are static. That is, the representation of each word is fixed regardless of the particular context in which it appears. The second generation of encoders are sequential models that are sensitive to the word sequence order. The Long Short Term Memory (LSTM) model (Hochreiter and Schmidhuber 1997) and its variants introduced a *memory cell* that is able to preserve a state over long periods of time. Several authors presented unsupervised, pre-trained, bidirectional LSTM–based encoders, such as CoVe (McCann et al. 2017) and ELMo (Peters et al. 2018), which can improve the performance on a wide variety of common NLP tasks. Next came graph-based approaches, which use a fully connected graph to model the relation of every two words within the input text. A successful implementation of this idea is the Transformer (Vaswani et al. 2017). In the Transformer, the neural network is pre-trained over a large text corpus under different self-supervised tasks such as Masked Language Modeling (MLM), Sequence to Sequence (Seq2Seq), Permuted Language Modeling (PLM), Denoising Autoencoder (DAE), Constrastive Learning (CTL), Replaced Token Detection (TRD), Next Sentence Prediction (NSP), and Sentence Order Prediction (SOP). Combining some of these tasks has also been proposed, such as BERT (Devlin et al. 2019), Transformer-XL, ALBERT, XLNet (Yang et al. 2019), ELECTRA (Clark et al. 2020), BART (Lewis et al. 2019), or GPT (Brown et al. 2020). Transformer-based Neural Language Models are able to solve tasks quite accurately from a limited set of training samples (via fine-tuning). Moreover, as Language Models, they have great predictive power over word strings (Radford et al. 2019; Brown et al. 2020).

However, it has been shown in the literature that contextual models do not always maintain the isometry with respect to the semantic similarity of words utterances. They concentrate the representation of words in hypercones of multidimensional space. This phenomenon has been called the **representation degradation problem** (Ethayarajh 2019; Gao et al. 2019; Li et al. 2020; Wu et al. 2020; Cai et al. 2021). The result is that contextual models are very predictive as Language Models, but not very effective in terms of text representation within a semantic space. In other words, although the neural network is able to predict words from the previous sequence and classification labels, the embedding space in which the texts are represented is not coherent with their meanings. According to different authors, such as Gao et al. (2019) and Demeter, Kimmel, and Downey (2020), this is due to an effect of infrequent words in the soft max optimization function. We will see that this limitation has an effect on our experiments. In addition, previous experiments show that the effect of the representation degradation problem is stronger in higher net layers where the word representations are more contextualized (Ethayarajh 2019). However, solving this problem is not the focus of the article and we leave its analysis for future work.

Some approaches, such as Sentence-BERT (Reimers and Gurevych 2019) and the Universal Sentence Encoder (Cer et al. 2018), address this problem by training the network over sentence pairs as a similarity classification task. However, the effectiveness

of these models may decline when the compared texts have different characteristics from the text units on which they have been trained. In particular, Raffel et al. (2020) found that supervised transfer learning from multiple tasks does not outperform unsupervised pre-training. Yogatama et al. (2019) conducted an extensive empirical investigation to evaluate state-of-the-art Natural Language Understanding models through a series of experiments that assess the task-independence of the knowledge being acquired by the learning process. They concluded that the performance is sensitive to the election of the supervised training task. We confirmed this phenomenon in our own experiments.

Recent experiments suggest through probes that Neural Language Models do not capture the systematic nature of language (Talmor et al. 2020; Pimentel et al. 2020; Goodwin, Sinha, and O'Donnell 2020; Hupkes et al. 2020; Bender and Koller 2020). When probing, a researcher chooses a linguistic task and trains a supervised model to predict annotations in that linguistic task from the network's learned representations. One of the most meticulous experiments was performed by Hupkes et al. (2020). This team developed five behavioral tests in order to analyze whether neural networks are able to generalize compositional aspects. The authors found that, for the majority of these tests, Recurrent, Convolution-based, and Transformer models fail. In summary, although tremendously powerful, Neural Language Models alone cannot represent previously unseen textual information by composition.

2.3 Compositional Distributional Semantic Models

In the Compositional Distributional approach, texts longer than one word are represented by means of a composition function that combines distributional representations of linguistic units. A large body of literature has shown that the sum or global average of word embeddings is very effective, often outperforming more sophisticated methods (Mitchell and Lapata 2010; Boleda 2020; Lenci 2018; Blacoe and Lapata 2012; Perone, Silveira, and Paula 2018; Baroni and Lenci 2010; Rimell et al. 2016; Czarnowska, Emerson, and Copestake 2019; Wieting and Gimpel 2018; Ethayarajh 2018). An intrinsic limitation of the additive approaches is that word order is not considered, since these composition functions are associative and, in addition, some experiments suggest that their effectiveness degrades with sentence length (Polajnar, Rimell, and Clark 2014). The tensor product has also been studied as a composition function (Clark and Pulman 2007); its main disadvantage is that the space complexity grows exponentially as more constituents are composed together.

As mentioned in the previous section, the Principle of Compositionality states that the composite meaning depends on the meaning of constituents and their syntactic relationships. In order to capture the information of linguistic structures, some researchers have trained an additive composition function on the basis of expression equivalences (Zanzotto et al. 2010), sentence similarity and paraphrasing tasks (Mitchell and Lapata 2008; Wieting et al. 2015), or sentiment labels in movie reviews (Socher et al. 2012). Their main limitations are that the composite structures are biased by the nature of the training corpora and are difficult to scale on more complex linguistic structures.

In another line of work, some authors proposed adding a symbolic layer on top of the distributional word representation. More specifically, Coecke, Sadrzadeh, and Clark (2010) proposed a composition function relying on the algebra of pre-groups and checked empirically the preservation of the vector dot product as an approach to similarity, although this property was not formally derived. Its main drawback is scalability

(Zhang 2014). It has been applied to relational words (Grefenstette and Sadrzadeh 2011); simple phrases (Kartsaklis, Sadrzadeh, and Pulman 2012); and pronouns, prepositions, and adjective phrases (Zhang 2014). Smolensky et al. (2016) presented an initial work on mapping inference in predicate logic based on tensor product representations. This approach is limited by the need for a previous mapping between phrases or sentences and logical propositions.

Finally, other authors have grounded their compositional approach on Information Theory notions. Arora, Liang, and Ma (2017) proved that applying IDF (word specificity[3]) weighted sum of vector plus Singular Value Decomposition achieves competitive results regarding some sequential models such as LSTM. Zhelezniak, Savkov, and Hammerla (2020) represented sentences as a sequence of random vectors and experimented with their Mutual Information approach as a sentence similarity measure. A common limitation of both approaches is that they do not consider the syntactic structure of the sentence.

In the context of the state of the art, we can situate our work as follows. We propose a framework for Compositional Distributional representation and study the formal desirable properties of embedding, composition, and similarity functions. Our work focuses on the problem of representation itself, taking as an objective the correspondence between distributional representations and meanings.

3. Theoretical Framework

In this section, we describe the proposed theoretical framework. First, we establish a geometric interpretation of Distributional Semantics and its connection with Information Theory. Next, we formally define the notion of Information Theory–based Compositional Distributional Semantics and its formal properties.

3.1 Geometrical Interpretation of Distributional Semantics

First, our theoretical framework is built on a phenomenon that has been formally justified and observed repeatedly in the literature: *There exists a correspondence between the vector norm and the specificity or IC of the represented utterance*. Formally, according to the analysis by Levy and Goldberg (2014) and Arora et al. (2016), the dot product of SGNS embedding approximates the Pointwise Mutual Information (PMI) between two words. With $\pi(w)$ being the embedding of the word w:

$$\langle \pi(w), \pi(w') \rangle \propto \mathrm{PMI}(w, w') = \log\left(\frac{P(w, w')}{P(w') \cdot P(w')}\right)$$

This implies that there exists a correspondence between the vector norm and the IC of represented utterances according to Shannon's Information Theory:

$$\mathrm{IC}(w) = -\log(P(w)) = -\log\left(\frac{P(w, w)}{P(w) \cdot P(w)}\right) = \mathrm{PMI}(w, w) \simeq \langle \pi(w), \pi(w) \rangle = \|\pi(w)\|^2$$

3 Note that $IDF(t) = \left(\log(\frac{N}{|\{d \in D : t \in d\}|})\right)$, where D represents the document collection and N represents its size, has a direct correspondence with its specificity of Information Content $IC(t) = -\log(P(t))$ according to the Shannon's Information Theory.

In addition, Gao et al. (2019) proved that under some assumptions, the optimal embeddings of infrequent tokens in Transformer Language Models can be extremely far away from the origin. Li et al. (2020, page 9121) observed empirically that "high-frequency words are all close to the origin, while low-frequency words are far away from the origin" in Transformer language models.

On the other hand, this phenomenon makes sense from the point of view of text representation. The absence of textual information should be a singular point in the representation space (origin of coordinates). This point should be equidistant to any text with a fixed amount of information. In other words, the set of possible texts with a fixed amount of information should form a sphere around the empty information point. In short, it makes sense that the representations are distributed around the origin of coordinates at a distance proportional to the amount of information they contain.

Figure 2 illustrates this idea. The longer a word sequence is, the lower its probability, increasing its specificity and, therefore, its IC. Consequently, as we remove words from an utterance, its embedding approaches the origin of coordinates (vector norm tends to zero). At the end, the probability of the empty word set is maximal, and therefore, its IC is minimal.

Another consequence of this geometrical interpretation of Distributional Semantic representation is that *the meaning in the pragmatic context of any utterance must be represented as an infinite norm vector* (human thinking in Figure 2). Assuming that the meaning in the pragmatic context of any utterance is extremely specific (time, place, actors, physic context, open-ended world knowledge, etc.), then we can assert that its probability in the semantic space tends to zero and therefore its IC according to Shannon's Theory is infinite. The representation is therefore theoretically an infinite vector and its proximity depends exclusively on its angular distance.

Because it is a continuous representation space, the embedding distances should approximate the semantic similarity of their utterances. However, we can consider two different notions of semantic similarity. On the one hand, *pragmatic semantic similarity* refers to the meaning proximity of two utterances taking into account the context in which they are used. In our geometric space, since the pragmatic meaning corresponds with infinite norm vectors, this would translate to the angular distance between infinite norm vectors. In fact, cosine similarity is the standard measure in distributional representations.

On the other hand, *literal similarity* takes into account the amount of information provided by the utterance. For example, although *"hello"* may be equivalent to *"hello*

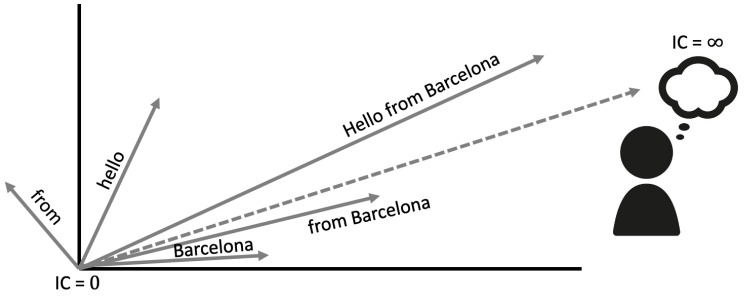

Figure 2
Geometric interpretation of text embedding based on Information Theory.

from Barcelona" in a certain context of use, there is a difference in terms of literal similarity, since *"hello"* is under-specified regarding *"hello from Barcelona."* That is, the space of possible meanings in the context of *"hello"* is broader and therefore its literal IC is lower. In empirical terms, literal utterance similarity represents the expected similarity given an annotator who is shown the two utterances without any information about their pragmatic context.

We can summarize the geometrical interpretation of embedding models as follows: *The direction of an embedding represents the pragmatic meaning, and the vector norm of embedding represents how much information the literal utterance provides about its meaning in the pragmatic context*. Returning to the example in Figure 2, the meaning of *"Barcelona"* depends on its pragmatic context. As we add words (*"from Barcelona," "hello, from Barcelona"*), the meaning in the pragmatic context of what the sender means to express becomes more precisely defined. That is, as we add context information to the expression when adding words, the embedding norm will become progressively longer and its semantic orientation more precise. Only in infinity do we arrive at the meaning in the pragmatic context. In this paper, we identify the properties that embedding, composition, and similarity functions should fulfill based on this geometric interpretation.

3.2 Information Theory–based Compositional Distributional Semantics

We formalize the notion of Information Theory–based Compositional Distributional Semantics as a tuple of three functions, namely: *embedding*, *composition*, and *similarity*. This framework synthesizes the distributional and compositional paradigms; the hypothesis underlying the notion of ICDS is that *there are minimal linguistic units whose semantics are determined by their use and whose amount of information is determined by their specificity. On the other hand, the systematicity of language can be captured by compositional mechanisms while preserving the amount of information of the composite utterance*.

Figure 3 illustrates the ICDS framework. The starting point in this example is a set of three words: *"yellow," "star,"* and *"galaxy."* The semantics of these atoms are inferred on the basis of their use or distributional properties, and not by composition. In our

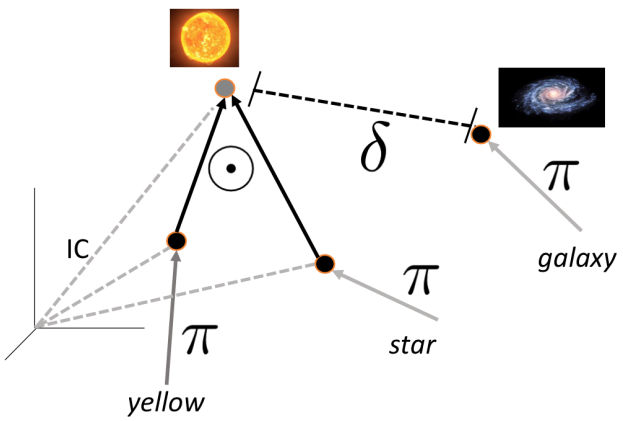

Figure 3
Embedding (π), Information Content (IC, vector norm), Composition (⊙), and Similarity (δ) functions in an Information Theory–based Compositional Distributional Semantics system.

framework, we assume that there exists an embedding function $\pi : S \longrightarrow \mathbb{R}^n$, which returns a vector representation for each basic linguistic unit x in the space S of basic linguistic units. Note that these semantic atoms are words but they could be multiword terms such as Named Entities or compound terms in general. For instance, it is not possible to infer the semantics of *"John Smith"* as a semantic composition of *"John"* and *"Smith"* and the semantics must be inferred from its use as a distinct linguistic unit. The same is true for some common terms; for example, the concept of *"black hole"* is not the semantic composition of the concepts of *"hole"* and *"black."* In our framework, we will refer to these atoms as **basic linguistic units**. The framework also assumes that there exists a composition function $\odot$ that returns a new embedding given two representations and a similarity function δ that estimates the embedding meaning proximity. The particularity of ICDS is that both the composition and the similarity functions must be consistent with the embedding Information Content (vector norms).

In Section 2 we discussed four aspects that a semantic representation model should cover as desirable properties. The embedding function π captures USAGE CONTEXT and INFORMATION MEASURABILITY. The composition function $\odot$ addresses SYSTEMATICITY and, whenever it satisfies certain properties, it maintains the INFORMATION MEASURABILITY coherence of composite representations. Finally, the semantic similarity function δ defines the continuous space of semantic representations.

The key point is that this formal framework captures the duality between Compositional and Distributional Semantics. The semantics of the basic linguistic units modeled by means of the embedding function π is determined by the textual context in which they appear, but, in turn, the semantics of complex structures is determined by the composition function $\odot$ and the way in which words are combined.

4. Formal Definition and Properties

In this article, we define the notion of Information Theory–based Compositional Distributional Semantics as follows:

Definition 1 (ICDS)
An Information Theory–based Compositional Distributional Semantics representation is a tuple $(\pi, \delta, \odot)$ *with an embedding, semantic similarity, and composition function. S being the space of basic linguistic units:*

$$\pi : S \longrightarrow \mathbb{R}^n,$$
$$\delta : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R},$$
$$\odot : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

In the following sections we formalize a set of formal properties for each of the components of the above definition of ICDS. We cannot be sure that there are no other possible properties that can be added in future work. However, based on our review of related work, what we can be sure is that this set of nine properties is sufficient to capture the particularities and differences among existing embedding functions.

4.1 Embedding Function Properties

The first two properties in Definition 1 affect the embedding function. In the following, we will denote the atomic linguistic units (words, chunks) as x, y, or z.

Property 1 (INFORMATION MEASURABILITY)
x being a linguistic unit, the vector norm of its embedding approximates the Information Content of x.

$$\|\pi(x)\| \simeq IC(x) = -\log(P(x))$$

INFORMATION MEASURABILITY formalizes what has already been observed in the experiments developed throughout the literature. That is, the correspondence between embedding vector norm and word specificity or frequency (see Section 3.1). It states that the embedding of the word sequence allows for estimating the probability of observing the corresponding word sequence, which is projected within representation space by π on the vector terminal point defined by the embedding vector. This property establishes an explicit link between text embedding and Information Theory.

The consequence is that stopwords and other frequent tokens (with low expected Information Content) are represented near the origin while infrequent words (with high expected Information Content) are projected further away from the origin of coordinates. In general, both Static and Contextual Neural Distributional Models comply with this property (Arora et al. 2016; Levy and Goldberg 2014; Gao et al. 2019; Li et al. 2020).

Property 2 (ANGULAR ISOMETRY)
There must exist isometry between the angular position of embeddings and the expected similarity of utterances according to humans. Being $|\pi(x)| = |\pi(y)|$*:*

$$\cos(\pi(x), \pi(y)) \propto \mathbb{E}(\text{SIM}(x, y))$$

where $\mathbb{E}(\text{SIM}(x, y))$ represents the expected similarity according to humans, which corresponds with the pragmatic context of the utterances. We assume the hypothetical case of a representative set of humans who are presented with the utterances, and who are asked about their semantic similarity.

ANGULAR ISOMETRY states the correspondence between the embedding direction and the (pragmatic) meaning. This property formalizes the idea of desirable semantic isometry, that is, it establishes a correspondence between the representation of texts and their semantic similarity given a fixed degree of specificity (Information Content). Given two embeddings with similar length (equally informative according to the analysis in Section 3.1), the expected similarity between represented utterances according to human annotators should be correlated with the angle these vectors maintain.

4.2 Composition Function Properties

The next three properties concern the composition function $\odot$. Unfortunately, it is not possible for the composition function to predict the probability (i.e., IC) of a compound expression ("yellow star") given the probability of its components (P(*"yellow"*) and P(*"star"*)) and its syntactic relationship. This is in fact an inherent limitation of Language Models in terms of compositionality (see Section 2.1). However, we can state constraints for the IC of the composite expression. These boundaries define three formal constraints illustrated in Figure 4. The black and gray dots represent the component and compound representations, respectively.

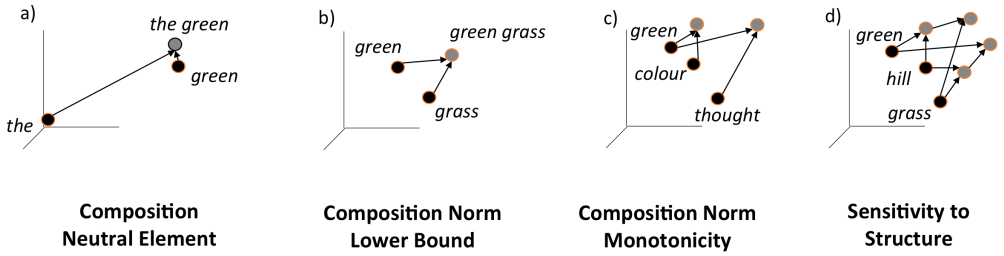

Figure 4
Illustration of composition function properties.

Property 3 (COMPOSITION NEUTRAL ELEMENT)
Null information components (zero norm) do not affect the composition.

$$\|\vec{v}_2\| = 0 \implies \|\vec{v}_1 \odot \vec{v}_2\| = \|\vec{v}_1\|$$

In principle, adding empty information should not affect the composition. It follows from COMPOSITION NEUTRAL ELEMENT that when a embedding with (nearly) empty information is composed with another given embedding, the resulting composite embedding will be very close to the given embedding. For instance, *"green"* is semantically similar to *"the green,"* assuming that *"the"* is not informative (Figure 4a).

Property 4 (COMPOSITION NORM LOWERBOUND)
The vector norm of the composite embedding is higher than or equal to the norm of each component; that is, the composition never reduces the Information Content.

$$\|\vec{v}_1 \odot \vec{v}_2\| \geq \|\vec{v}_1\| \qquad \|\vec{v}_1 \odot \vec{v}_2\| \geq \|\vec{v}_2\|$$

COMPOSITION NORM LOWERBOUND states that adding textual information should not reduce the final amount of information. That is, the IC (i.e., vector norm, see Section 3.1) of an utterance embedding increases when adding words. For instance, *"green grass"* is more informative than *"green"* and *"grass"* (Figure 4b).

Property 5 (COMPOSITION NORM MONOTONICITY)
The norm of the composite vector is monotonic with respect to the angle between the compound vectors:

$$\left.\begin{array}{l} \|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\| \\ \cos(\vec{v}_1, \vec{v}_2) > \cos(\vec{v}_1, \vec{v}_3) \end{array}\right\} \implies \|\vec{v}_1 \odot \vec{v}_2\| < \|\vec{v}_1 \odot \vec{v}_3\|$$

COMPOSITION NORM MONOTONICITY states that the more semantically dissimilar the components are (greater angular distance), the more information the composite embedding contains. This property responds to the intuition that combining two semantically distant utterances should produce specific and infrequent utterances, that is, with a high information content associated with it. For instance, *"green"* and *"colour"* are semantically close and their representation vectors should therefore point to nearby

areas in the infinite (Figure 4c). That is to say that *"green"* and *"colour"* have an angular distance that is less than that of two semantically unrelated words, such as *"green"* and *"thought."* The compound term *"green thought"* is therefore more specific, less likely, and more informative than *"green colour"*; and its vector norm should therefore be larger. In other words, under fixed single vector norms (equal IC), the semantic orientation similarity of components decreases the Information Content of the composition.

In addition to these three properties, there is another aspect to consider in ICDS. Compositionality requires sensitivity to the linguistic structure.

Property 6 (SENSITIVITY TO STRUCTURE)
Given the three embeddings $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ with an equal norm and angularly equidistant, their composition is not associative:

$$\left.\begin{array}{l} \|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\| > 0 \\ \cos(\vec{v}_1, \vec{v}_2) = \cos(\vec{v}_1, \vec{v}_3) = \cos(\vec{v}_2, \vec{v}_3) > 0 \end{array}\right\} \Longrightarrow (\vec{v}_1 \odot \vec{v}_2) \odot \vec{v}_3 \neq \vec{v}_1 \odot (\vec{v}_2 \odot \vec{v}_3)$$

Let us explain the motivation for SENSITIVITY TO STRUCTURE. The Principle of Compositionality assumes that linguistic units are progressively grouped into more complex semantic units, according to the structure of the language (Figure 4d). That is, the composition function must be applied on the representations according to how the words are considered to be structured. For example, assuming constituents, the composition function on the sentence *"This house is very big"* is applied following the brackets: *"((This house) is (very big))"*.

Unfortunately, we cannot define formal geometric properties that ensure that the distributional model captures the semantics of composition. However, a basic requirement is that the final representation is at least sensitive to the way in which the language is structured. That is, different ways of grouping words in the composition function should produce different representations. More formally, the vector composition function should not be associative.

4.3 Similarity Function Properties

The next three properties concern the similarity function δ (see Figure 5).

Property 7 (ANGULAR DISTANCE SIMILARITY MONOTONICITY)
Under equal vector norm (equal IC), similarity is a monotonic decreasing function with regard to the angular distance (and semantic orientation proximity).

$$\left.\begin{array}{l} \cos(\vec{v}_1, \vec{v}_2) > \cos(\vec{v}_1, \vec{v}_3) \\ \|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\| > 0 \end{array}\right\} \Longrightarrow \delta(\vec{v}_1, \vec{v}_2) > \delta(\vec{v}_1, \vec{v}_3)$$

ANGULAR DISTANCE SIMILARITY MONOTONICITY expresses the monotonicity of similarity regarding the semantic orientation of text representation vectors. Note that ANGULAR ISOMETRY (defined above) states the correspondence between the semantic similarity of text units and their embedding angle. ANGULAR DISTANCE SIMILARITY MONOTONICITY ensures that the similarity function δ is coherent with this principle. More formally, similarity between two representation vectors increases if the cosine of the angle formed by the vectors increases, that is, if the angle is reduced (Figure 5a).

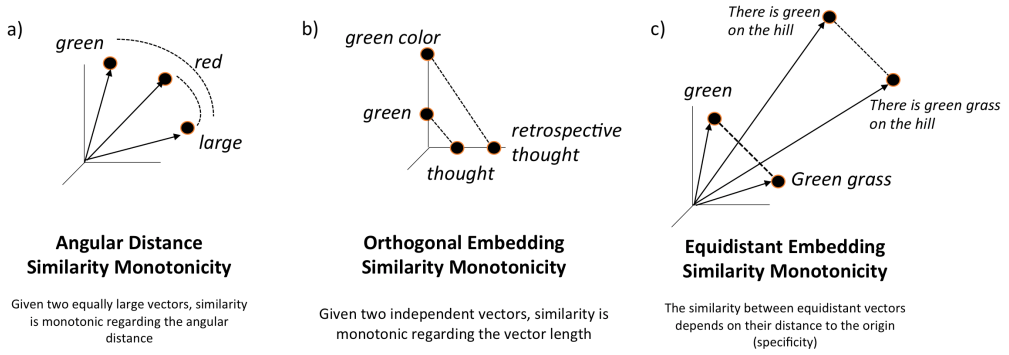

Figure 5
Illustration of the preconditions of the properties ANGULAR DISTANCE SIMILARITY MONOTONICITY, ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY, and EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY.

Property 8 (ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY)
Given orthogonal embeddings, the larger their norm is (their specificity), the less similar will be:

$$\left.\begin{array}{l}\cos(\vec{v}_1, \vec{v}_2) = \cos(\vec{v}_3, \vec{v}_4) = 0 \\ \|\vec{v}_1\| < \|\vec{v}_2\|, \|\vec{v}_3\| < \|\vec{v}_4\|\end{array}\right\} \Longrightarrow \delta(\vec{v}_1, \vec{v}_2) > \delta(\vec{v}_3, \vec{v}_4)$$

ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY expresses the decreasing monotonicity of similarity regarding the IC (i.e., vector norm) of components. That is, given two semantically unrelated utterances, the more they contain information, the more they are semantically dissimilar. Let us consider two semantically distant words such as *"green"* and *"thought,"* and two semantically distant utterances such as "green colour" and "retrospective thought" (Figure 5b). The more information we add to the semantically distant utterances, the lower their similarity should be.

Property 9 (EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY)
Given two embedding pairs $(\vec{v}_1, \vec{v}'_1)$ and $(\vec{v}_2, \vec{v}'_2)$, *and with* $\vec{\epsilon}$ *being a vector that represents their equidistance, then:*

$$\left.\begin{array}{l}\vec{v}'_1 = \vec{v}_1 + \vec{\epsilon},\ \ \vec{v}'_2 = \vec{v}_2 + \vec{\epsilon} \\ \|\vec{v}_2\| > \|\vec{v}_1\| >> \|\vec{\epsilon}\|\end{array}\right\} \Longrightarrow \delta(\vec{v}_2, \vec{v}'_2) > \delta(\vec{v}_1, \vec{v}'_1)$$

EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY captures the idea that given two utterances that are close to each other, their specificity gives them similarity. Figure 5c reflects this idea. The utterances *"green"* and *"green grass"* are not necessarily similar, given that the term *"green"* could refer to any kind of object. However, their similarity could be higher in a more specific context, for instance, *"the top of the mountain is green"* vs. *"the top of the mountain is green grass."* Under our geometrical interpretation, the further away the related embeddings are from the origin (more specificity and content), the more they are semantically similar. More formally, under equal Euclidean distance, semantic similarity grows with the embedding vector norm.

5. Function Analysis and Generalization

In this section we analyze existing embedding, composition, and similarity functions (the tuple that defines an ICDS) on the basis of the formal properties stated in our formal framework. In addition, we define novel parameterizable functions that generalize existing approaches for composing. We also introduce the linguistic structures that will be considered in our experiments.

5.1 Embedding and Information Content Functions

According to previous studies, both static (SGNS, GloVe) and contextual (BERT, GPT) embedding models keep the correspondence between the embedding vector norm and the specificity of the represented linguistic unit (see Section 3.1). Therefore, the vector norm can be used as Information Content function, satisfying INFORMATION MEASURABILITY.

Regarding ANGULAR ISOMETRY, previous work demonstrates the correspondence between the angular distance and the semantic similarity of words in static embedding models such as SGNS and GloVe. At a theoretical level, the dot product of Skip-gram with Negative-Sampling embedding approximates the PMI of words (Levy and Goldberg 2014; Arora et al. 2016).

$$\langle \pi(x), \pi(y) \rangle \propto \mathrm{PMI}(x, y) = \log \left(\frac{P(x, y)}{P(x) \cdot P(y)} \right)$$

Note that under the distributional hypothesis (*"similar contexts imply similar meaning"*), PMI should approximate the semantic similarity between utterances. That is, a high PMI between two utterances implies a high co-occurrence ($P(x, y)$) with respect to their likelihood in isolation ($P(x)$ and $P(y)$). Consequently, two infrequent terms that co-occur will have a high PMI, while a frequent term (e.g., a determiner) will have a low PMI with any other term.

The main drawback of static embedding functions is that the occurrence's local context is not considered. In contrast, Contextual Neural Language Models (e.g., BERT, GPT) take into account the textual context of language units when embedding them (see Section 2). However, due to the representation degeneration problem (see Section 2.2), contextual models suffer from limitations in terms of ANGULAR ISOMETRY (Property 2).

5.2 Composition Functions

In this section we study the properties of basic composition functions such as summation or averaging. The global average, namely, the direct average of all embedding vectors that make up an utterance, and sum have been found to be strong baselines for compositional tasks (Boleda 2020; Lenci 2018; Blacoe and Lapata 2012; Perone, Silveira, and Paula 2018; Baroni and Lenci 2010; Rimell et al. 2016; Czarnowska, Emerson, and Copestake 2019; Wieting and Gimpel 2018; Ethayarajh 2018).

In addition, in order to satisfy the formal properties, we propose a generalized composition function based on Information Theory that can be instantiated in different measures according to the chosen parameters.

Table 2
Formal constraints analysis for embedding, composition, and similarity functions.

	Angular Isometry	Inf. Measurability		Comp. Neutral Element	Comp. Norm Lowerbound	Comp. Norm Mon.	Sensitivity to Structure		Angular Dist. Sim. Mon.	Orth. Emb. Sim. Mon.	Equid. Emb. Sim. Mon.
Static	✔	✔	Global average	✗	✗	✗	✗	Euclidean	✔	✔	✗
(SGNS, GloVe)			$\mathtt{F}_{\mathtt{Sum}}$ $(\lambda = 1, \mu = -2)$	✔	✗	✗	✗	Cosine	✔	✗	✔
Contextual	✗	✔	$\mathtt{F}_{\mathtt{Avg}}$ $(\lambda = \frac{1}{4}, \mu = -\frac{1}{2})$	✗	✗	✗	✔	Dot prod.	✔	✗	✔
(BERT,GPT)			$\mathtt{F}_{\mathtt{Ind}}$ $(\lambda = 1, \mu = 0)$	✔	✔	✗	✔	ICM	✔	✔	✔
			$\mathtt{F}_{\mathtt{Joint}}$ $(\lambda = 1, \mu = 1)$	✔	✗	✔	✔				
			$\mathtt{F}_{\mathtt{Inf}}$	✔	✔	✔	✔				

Definition 2 (Generalized Composition Function)
The generalized composition function $\mathtt{F}_{\lambda,\mu} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ *is defined as:*

$$\mathtt{F}_{\lambda,\mu}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|} \cdot \sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}$$

The first component on the right-hand side of the expression is the unit vector of their sum and determines the vector direction. The second component determines the vector norm (magnitude), which depends on the norm of single vectors and their inner product. This function generalizes the most common composition functions in the literature. In particular, it leads to the sum of vectors ($\mathtt{F}_{\mathtt{Sum}}$) when $\lambda = 1$ and $\mu = -2$, and to the pairwise average vector[4] ($\mathtt{F}_{\mathtt{Avg}}$) when $\lambda = \frac{1}{4}$ and $\mu = -\frac{1}{2}$.

$\mathtt{F}_{\mathtt{Sum}}$ and global average do not comply with Composition Lowerbound (Property 4, see Table 2). For instance, two vectors pointing to opposite directions can cancel each other. Furthermore, they do not comply with Composition Norm Monotonicity (Property 5). In fact, the more distant in orientation the original vectors are, the shorter the length of the composite vector. In addition, the sum and global average are the only non-associative functions in this set. Therefore, they also fail to satisfy Sensitivity to Structure (Property 6). The four remaining functions are sensitive to the text structure. Global average does not comply with COMPOSITION NEUTRAL ELEMENT either.

4 The difference between pairwise and global averaging lies on the composition procedure for utterances of N components with N being bigger than 2. The former applies pairwise compositions until the ROOT node is reached (2 components are considered at each step) while the latter composes all terms in a single averaging step.

On the other hand, pairwise average vector ($\mathtt{F}_{\mathtt{Avg}}$) is sensitive to the addition of null information (e.g., $avg(4,0) = 2 < 4$). Therefore, it does not satisfy COMPOSITION NEUTRAL ELEMENT (Property 3). It also fails to satisfy COMPOSITION NORM MONOTONICITY (Property 5).

In addition to the basic composition function, $\mathtt{F}_{\lambda,\mu}$ also generalizes certain Information Theory–based interpretable functions. Let $\mathtt{F}_{\mathtt{Ind}}$ denote the particularization $\lambda = 1$ and $\mu = 0$. In $\mathtt{F}_{\mathtt{Ind}}$, the IC (vector norm) of the composite vector depends exclusively on the IC of single vectors. In other words, $\mathtt{F}_{\mathtt{Ind}}$ assumes that the combined linguistic forms are statistically independent, and, therefore, that the Information Content in the composition is additive. This means that the IC of the composition does not depend on the degree of statistical co-occurrence of the language forms being combined:

$$\|\mathrm{F}_{Ind}(\pi(x),\pi(y))\|^2 = \|\pi(x)\|^2 + \|\pi(y)\|^2 \simeq \mathrm{IC}(x) + \mathrm{IC}(y)$$

$\mathtt{F}_{\mathtt{Ind}}$ does not satisfy COMPOSITION NORM MONOTONICITY since the angular distance does not affect the composite embedding.

Let $\mathtt{F}_{\mathtt{Joint}}$ denote the particularization $\lambda = 1$ and $\mu = 1$. Assuming correspondence between PMI and the dot product, in $\mathtt{F}_{\mathtt{Joint}}$, the Information Content of the composite vector is the joint Information Content of the components.[5] This means that the IC of the composition depends exclusively on the degree of statistical co-occurrence ($P(x,y)$) of the language forms being combined. Low co-occurrence words will produce more information when combined than high co-occurrence words:

$$\|\mathrm{F}_{Joint}(\pi(x),\pi(y))\|^2 = \|\pi(x)\|^2 + \|\pi(y)\|^2 - \langle\pi(x),\pi(y)\rangle \simeq -\log(P(x,y)) = \mathrm{IC}(x,y)$$

Because the composite embedding corresponds to the joint IC in $\mathtt{F}_{\mathtt{Joint}}$, it does not comply with COMPOSITION NORM LOWERBOUND (Property 4).

None of the previous composition functions satisfy all properties simultaneously. The following theorem establishes the range of values of λ and μ for which all formal properties are satisfied. This theorem implies that none of the previous functions ($\mathtt{F}_{\mathtt{Joint}}$, $\mathtt{F}_{\mathtt{Ind}}$, $\mathtt{F}_{\mathtt{Sum}}$, and $\mathtt{F}_{\mathtt{Avg}}$) satisfies the three composition properties simultaneously. However, the functions $\mathtt{F}_{\mathtt{Joint}}$ and $\mathtt{F}_{\mathtt{Ind}}$ are just at the boundaries (see proof in Section A.1):

Theorem 1
The generalized composition function satisfies COMPOSITION NEUTRAL ELEMENT *(Property 3)*, COMPOSITION NORM LOWERBOUND *(Property 4)*, COMPOSITION NORM MONOTONICITY *(Property 5)*, and SENSITIVITY TO STRUCTURE *(Property 6) if and only if* $\lambda = 1$ *and* $\mu \in \left(0, \frac{\min(\|\vec{v}_1\|,\|\vec{v}_2\|)}{\max(\|\vec{v}_1\|,\|\vec{v}_2\|)}\right]$.

Let $\mathtt{F}_{\mathtt{Inf}}$ denote to the particularization $\lambda = 1$ and $\mu = \frac{min(\|\vec{v}_1\|,\|\vec{v}_2\|)}{max(\|\vec{v}_1\|,\|\vec{v}_2\|)}$.The motivation for this parameter instantiation is, first, that it falls within the theoretical range in which the properties of composition are satisfied. Second, it satisfies the property that

5 Note that: $\|\pi(x)\|^2 + \|\pi(y)\|^2 - \langle\pi(x),\pi(y)\rangle \simeq -\log(P(x)) - \log(P(y)) - \log\left(\frac{P(x,y)}{P(x)\cdot P(y)}\right) = -\log(P(x,y))$

composing two vectors with the same direction results in the longer one. This means that adding redundant information does not affect the original embedding. Therefore it does not increase the amount of information. For instance, two repeated sentences have the same meaning as one of them. More formally:

$$\cos(\vec{v}_1, \vec{v}_2) = 1 \wedge \|\vec{v}_1\| > \|\vec{v}_2\| \Longrightarrow \mathtt{F}_{inf}(\vec{v}_1, \vec{v}_2) = \vec{v}_1$$

Note that this parameter fixing is somehow arbitrary. According to Theorem 1, a suitable composition function parameterization should be fixed between $\mu > 0$ ($\mathtt{F}_{\mathtt{Ind}}$ instantiates $\mu = 0$) and $\mu = \frac{\min(\|\vec{v}_1\|, \|\vec{v}_2\|)}{\max(\|\vec{v}_1\|, \|\vec{v}_2\|(\mathtt{F}_{Inf})}$. At one end ($\mathtt{F}_{\mathtt{Ind}}$, at the boundary but not included) the IC of the utterance *"hello hello"* doubles the IC of the word *"hello."* At the other end ($\mathtt{F}_{\mathtt{Inf}}$, at the boundary and included) both utterances are equivalent. Both functions represent the boundaries of the formal parameter range.

Thus, any other choice of parameter values within the formal range will give us a function where the IC of "hello hello" will be between the IC of *"hello"* and the double of this value. At the moment, we have no information beyond that contained in the embedding vectors, so no pragmatic aspect is considered in the composition. However, the theory leaves room for including this kind of information, so that the IC of *"hello hello"* would be larger than the IC of *"hello."* Since we do not currently have a criterion for what the IC difference should be between these two cases, for our $\mathtt{F}_{\mathtt{Inf}}$ function we are left with the value at the extreme end of the theoretical range, and it does not increase the amount of IC.

In conclusion, we can state that, unlike the most common composition functions in the literature, the proposed function $\mathtt{F}_{\mathtt{Inf}}$ is the only one that simultaneously satisfies all four properties.

5.3 Similarity Functions

The most popular vector similarity functions in the literature are Euclidean distance, cosine similarity, and the inner product. Table 2 summarizes the formal properties of these similarity functions.

Euclidean distance is monotonic regarding both ANGULAR DISTANCE SIMILARITY MONOTONICITY (Property 7) and ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY (Property 8). However, it is invariant with respect to the origin of coordinates. Therefore, it does not satisfy EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY (Property 9). On the other hand, cosine similarity focuses exclusively on angular distance, satisfying ANGULAR DISTANCE SIMILARITY MONOTONICITY (Property 7) but not ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY (Property 8). EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY (Property 9) is satisfied given that the cosine of two close points increases as they move away from the origin of coordinates if the Euclidean distance between them is maintained.

The third typical similarity approach is the inner product ($\langle \vec{v}_1, \vec{v}_2 \rangle = \|\vec{v}_1\| \|\vec{v}_2\| \cos(\vec{v}_1, \vec{v}_2)$), which has a correspondence with PMI in distributional models. The inner product is monotonic regarding the angle between vectors, and it is sensitive to vector norms. It lacks the property of ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY. The reason for that is that orthogonal vectors achieve zero similarity regardless of their norm.

As an alternative, the Information Contrast Model (ICM) (Amigó et al. 2020) is a generalization of the well-known PMI model, which adds three parameters α_1, α_2, and β. Then, x and y being two probabilistic events:

$$\mathrm{ICM}_{\alpha_1,\alpha_2,\beta}(x,y) = \log\left(\frac{P(x,y)^{\beta}}{P(x)^{\alpha_1}\cdot P(y)^{\alpha_2}}\right) = \alpha_1\mathrm{IC}(x) + \alpha_2\mathrm{IC}(y) - \beta\mathrm{IC}(x,y)$$

ICM ranges between $-\infty$ and ∞, and is equivalent to PMI when $\beta = 1$. It is also equivalent to the product of conditional probabilities when $\beta = 2$. ICM also generalizes the Linear Contrast Model proposed by Tversky (1977) under certain assumptions.

Assuming symmetricity ($\alpha_1 = \alpha_2 = 1$), assuming that the inner product approaches the PMI of the projected linguistic units,

$$\langle x,y\rangle \simeq PMI(x,y) \simeq \left(\frac{P(x,y)}{P(x)\cdot P(y)}\right) \simeq \mathrm{IC}(x) + \mathrm{IC}(y) - \mathrm{IC}(x,y)$$

and assuming INFORMATION MEASURABILITY ($\mathrm{IC}(x) \simeq \|\pi(x)\|$), then ICM can be rewritten as:

$$\begin{aligned} ICM^{V}_{\beta} &= \|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \beta(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \langle\vec{v}_1,\vec{v}_2\rangle) \\ &= (1-\beta)(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) + \beta\|\vec{v}_1\|\|\vec{v}_2\|\cos(\vec{v}_1,\vec{v}_2) \end{aligned}$$

We will refer to this similarity function as the **Vector-based ICM**. Interestingly, ICM^{V}_{β} is equivalent to the inner product when $\beta = 1$, and it is equivalent to the Euclidean distance when $\beta = 2$. In addition, ICM^{V}_{β} is the only one that satisfies the three similarity constraints (see Table 2 and formal proof in Section A.2):

Theorem 2
The Vector-based Information Contrast Model satisfies ANGULAR DISTANCE SIMILARITY MONOTONICITY *(Property 7)*, ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY *(Property 8), and* EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY *(Property 9)* whenever $\beta \in (1,2)$.

According to previous studies (Amigó and Gonzalo 2022), given a collection of document pairs, the optimal parameter can be estimated as the ratio between the average sum of Information Content and the average joint Information Content of all instance pairs for which similarity must be estimated:

$$\hat{\beta} = \frac{Avg\left(\mathrm{IC}(x) + \mathrm{IC}(y)\right)}{Avg\left(\mathrm{IC}(x,y)\right)} \tag{1}$$

In terms of the vector-based ICM:

$$\hat{\beta} = \frac{Avg\left(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2\right)}{Avg\left(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \langle\vec{v}_1,\vec{v}_2\rangle\right)} \tag{2}$$

The estimator $\hat{\beta}$ is designed to satisfy similarity formal properties, to give an average similarity value of zero (neutral), and also to give neutral similarity to document pairs with average single and joint Information Content.

5.4 Linguistic Structures

If the composition function satisfies SENSITIVITY TO STRUCTURE (all excepting $\mathtt{F}_{\mathtt{Sum}}$ and the global average), the linguistic structure and the order in which linguistic units are composed affects the composite embedding. Regarding the literature, past linguistic theories have discussed the sequential versus hierarchical processing of linguistic data (Frank, Bod, and Christiansen 2012).

For instance, in a sequential right-to-left manner, an example sentence would be structured as: *Just (Robert (showed (up (for (the (party)))))*. A left-to-right composition would be: *((((Just Robert) showed) up) for) the) party*. On the other hand, a hierarchical structure could be: *(Just Robert) ((showed up) (for (the party)))*.

Many structural constraints have been granted to the development of languages that may express and share intrinsic features of text semantics but also, some of these attributions to languages may have just appeared as a simple-to-explain aggregation of elements that do not relate directly to the Compositional Theorem of textual units. Therefore, an order must be specified to decide which child node is the next one to modify the parent via composition. In this work, we have explored a right-to-left and left-to-right order; but more complex approaches, such as a rule-based system considering the type of relations in dependency parsing or constituent types, could also be used. In this article, we consider three different linguistic structures. The first is the sequential composition of the data in which composition is based on a temporal order, assuming that the text is humanly understood either by the forward or backward composition/propagation of its components semantics. The second structure is constituency parsing, in which the role of sequential and component relations is more balanced. A third option is the dependency tree. In this case, sequential order plays a minor role in the composition and dependencies among elements are what drive the main structural process.

6. Experiments

The purpose of the experiments carried out in this article is to study to what extent the satisfaction of the described formal properties affects the effectiveness of meaning representation models. In this sense, the combination of the Word2Vec embedding function, the $\mathtt{F}_{\mathtt{Inf}}$ combining function, and the ICM similarity, which satisfy all formal properties, should equal or improve the rest of approaches. Our experiments have been sensitive enough to capture the positive effect of the properties ANGULAR ISOMETRY, COMPOSITION NEUTRAL ELEMENT, SENSITIVITY TO STRUCTURE, and EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY.

Because most of the composition functions are not associative, they are sensitive to the linguistic structure that guide the representation composition. That is, considering different linguistic structures lead to different word groupings in the composition process. In this article we compare the most common linguistic structures in NLP tools, namely, bag of words (no structure), sequential, constituent, and dependency parsing.

Our main findings are:

1. In the absence of supervision, contextual embedding functions (BERT and GPT-2) perform worse than non-contextual (static) word embedding functions (Word2Vec). Note that, according to the literature, the contextual embedding functions have limitations in terms of ANGULAR

ISOMETRY due to the representation degradation problem (Ethayarajh 2019; Gao et al. 2019; Li et al. 2020; Wu et al. 2020; Cai et al. 2021).

2. Composition functions based on Information Content ($\mathtt{F}_{\mathtt{Ind}}$, $\mathtt{F}_{\mathtt{Joint}}$, and $\mathtt{F}_{\mathtt{Inf}}$) perform better than traditional composition functions ($\mathtt{F}_{\mathtt{Sum}}$, $\mathtt{F}_{\mathtt{Avg}}$), suggesting that COMPOSITION NEUTRAL ELEMENT and SENSITIVITY TO STRUCTURE are relevant properties.

3. Using models trained on similarity tasks such as S-BERT improves non-supervised similarity measures on tasks similar to those on which they have been trained. However, the effectiveness decreases substantially when applied to data of a different nature. This corroborates the limitations of current contextual embedding models as meaning representation methods in non-supervised scenarios (ANGULAR ISOMETRY).

4. The ICM similarity function generalizes both the Euclidean distance and the dot product (see Section 5.3). Unlike these, ICM satisfies simultaneously ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY and EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY. The experiments show that ICM outperforms its components in isolation. However, the cosine similarity, which simply ignores the vector norms (ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY), achieves similar or even better results than ICM. Unfortunately, this suggests that our datasets are not sensitive to the effect of the IC (i.e., vector norm) of the compared utterances.

5. Although considering structures improves the results (SENSITIVITY TO STRUCTURE) with respect to bag of words-based approaches (sum, global average), we did not observe any substantial and consistent difference between linguistic structures (sequential, constituent, and dependency).

6.1 Evaluation Metric

Models of distributional representation have been evaluated in multiple ways throughout the literature. In many cases, extrinsic evaluation is applied, studying the effectiveness of the model for the performance of a particular task for which training data are available. However, this evaluation criterion may be biased by the task itself and also captures aspects that are not specific to the representation, such as behavior against training samples.

Other evaluation metrics focus on the characteristics of the model as a representation system. Some of them focus on the sensitivity of the model such as *non-conflation* or *Demonstration of Multifacetedness* (Wang et al. 2019). In view of the literature, it is not an aspect that discriminates too much between models.

Another common criterion is *analogy*, which states that the representation model must maintain a certain coherence in how the words relate to each other. That is, if there is a relationship between two words, for example, *France-Paris*, then from *Germany*, it should be possible to infer *Berlin*. Rogers, Drozd, and Li (2017) distinguish between four types of analogies: inflectional morphology, derivational morphology, lexicographic semantics, and encyclopedic. The Bigger Analogy Test Set contains 15 types of linguistic relationships (Gladkova and Drozd 2016). Analogy as an evaluation

criterion was referenced when word embedding models became popular (Mikolov et al. 2013) and in subsequent work (Baroni, Dinu, and Kruszewski 2014; Levy and Goldberg 2014; Schnabel et al. 2015; Rogers, Drozd, and Li 2017; Wang et al. 2019). Furthermore, analogy has been exploited in different NLP tasks such as word sense disambiguation, search, and so forth (Rogers, Drozd, and Li 2017). However, this assessment criterion is not appropriate in our context, since it focuses on lexical representation. Note that we are interested in assessing the embedding compositionality.

On the other hand, *contextuality* metrics quantify the extent to which the representation model captures the context in which words occur. Some of the measures proposed in the literature are the average cosine distance between representations of the same word in different contexts, or between the representation of a sentence and the set of its words. Ethayarajh (2019) studied the variance of principal components of the matrix formed by the embeddings of a word in different contexts. However, this evaluation criterion is also lexical-oriented.

Isotropy is another criterion for assessing representation. It is assumed that the distribution of embeddings in the space must have certain a priori characteristics, that is, the distribution of embeddings should look the same regardless of the reference point. Some measures that capture this aspect are the distribution of scalar products with the average vector (Mimno and Thompson 2017), the average dot product with unit vectors (Mu and Viswanath 2018), the average distance to their nearest neighbor of frequent versus infrequent terms (Li et al. 2020), or the average cosine distance between embeddings (Cai et al. 2021). Isotropy is related to the capacity of the representation model to use the full dimensional space. It has been shown to have some positive effect on the effectiveness of representation models, but in this article we will focus on less indirect aspects.

In our experiments, we will focus on *isometry*. That is, there must exist a similarity function over the embedding space that correlate with the proximity of meanings. Several authors directly compute the correlation between embedding cosine and lexical similarity annotated by humans (Gladkova and Drozd 2016; Qiu et al. 2018; Wang et al. 2019). Some available lexical similarity datasets are Wordsim-353 (Finkelstein et al. 2001), MEN (Bruni, Tran, and Baroni 2014a), SimLex-999 (Hill, Reichart, and Korhonen 2015), or SimVErb-3500 (Gerz et al. 2016). There are also some variants in which instead of graded similarity values, the embedding proximity is compared against synonyms or term categories (Baroni, Dinu, and Kruszewski 2014). For instance, Mu and Viswanath (2018) evaluated clustering algorithms over word representations and Cohenen et al. clustered representations according to their sense and quantified the discriminative power of distance to sense centroids (Mu and Viswanath 2018). Some word/concept datasets are Almuhareb (2006), ESSLLI 2008 Distributional Semantic Workshop shared task dataset (Baroni, Evert, and Lenci 2008), or the Batting test set (Baroni and Lenci 2010). Detecting an outlier in a term set is another isometry-based metric (Camacho-Collados and Navigli 2016). Another approach consists of measuring the component-wise correlation between word vectors from a word embedding model and manually constructed linguistic word vectors in the SemCor dataset (Miller et al. 1994), that is, Qvec (Tsvetkov et al. 2015).

But all the previous metrics are lexically oriented. Isometry can be extrapolated to sentence representation via semantic text similarity annotations (Zhang et al. 2020; Poerner, Waltinger, and Schütze 2020; Li et al. 2020; Yang et al. 2019). Note that testing isometry is distinct from using textual similarity as an extrinsic task-oriented metric. That is, a neural distributional model might not provide a match between representation and meaning spaces (isometry), and yet provide very good results in supervised textual

similarity prediction with training data. This is the case of BERT (Devlin et al. 2019). It should also be noted that in our experiments, the term isometry is not entirely accurate from a formal point of view, since both the semantic similarity itself and some similarity functions (e.g., PMI, ICM) do not necessarily fulfill the properties of a metric space (Amigó et al. 2020; Amigó and Gonzalo 2022). For simplicity, we will use this term in the article to refer to our evaluation criterion.

In this article we will focus on the notion of non-supervised isometry, evaluating the correspondence between embedding distances and meaning similarity of sentences and paragraphs. This evaluation methodology checks simultaneously the suitability of the embedding, composition, and similarity functions.

6.2 Datasets

We have selected different publicly available datasets that annotate the similarity scores of two pieces of text based on a certain target.

STS. The most popular semantic similarity dataset, STS (Cer et al. 2017), corresponds to the evaluation set from the corpus of English STS shared task data (2012–2017). It contains 1,102 pairs of sentences annotated with their meaning similarity ranging from 0 to 5, with 0 being two completely dissimilar sentences and 5 being two completely equivalent sentences. This dataset is composed of multiple subsets (**MSRpar**, **MSRvid**, **answer-answer**, **track5**, and **images**).

SICK. In general, the STS dataset is oriented to ambiguity and lexical semantic proximity rather than linguistic structures, and many sentence pairs consist of the substitutions of terms with synonyms. Marelli et al. (2014) developed SICK, a dataset created from STS specifically to test Compositional Distributional Semantic models given that the pairs of sentences encode compositional knowledge.

DEF2DEF. A common feature of the above two corpora is that their sentence pairs tend to have a high lexical overlap. In order to experiment on purely semantic proximity, we have developed DEF2DEF, where the texts compared consist of definitions of words. For this purpose, we take word pairs from the MEN Test Collection. MEN consists of 2,993 pairs of words annotated with their lexical-semantic similarity ranging from 0 to 50 (Bruni, Tran, and Baroni 2014b). We replace the words with their first definition in the WordNet set of synsets, which corresponds to the most frequent use of the word and assume that the semantic similarity between their respective definitions does not differ greatly from the original annotations. To corroborate this assumption, we have carried out a parallel annotation in which two human experts identify between two pairs of definitions, the pair that has the highest similarity between its definitions; 88.5% of the 200 annotated pairs matched the corresponding word similarities from the original MEN corpus. We used the following steps to sample comparison pairs from MEN corpus: (i) Define number of samples to annotate (200) and a relatedness difference threshold (20). (ii) Extract pairs of words from the MEN corpus and their relatedness score RS_1. (iii) For each original pair of words, randomly select another pair of words from the MEN corpus and their relatedness score RS_2 such that $abs(RS_1 - RS_2) \geq 20$. Note that repeated pairs are avoided at this step. (iv) The new gold standard becomes 1 if $RS_1 > RS_2$, else 0. Hypothetically, word-level comparisons should match results of definition level comparisons. The model performance in all previous datasets is

evaluated in our experiments via the Pearson correlation between the predicted and groundtruth similarities.

WORD2DEF. In all previous datasets, sentence pairs have a similar semantic specificity or Information Content; for instance, "short sentence" to "short sentence" or "paragraph" to "paragraph" in STS and SICK datasets. In contrast, the MEN corpus contains word pairs with similar semantic specificity. Following the methodology introduced by Kartsaklis, Sadrzadeh, and Pulman (2012), we have built WORD2DEF, a custom benchmark based on proximity between words and their definitions. A good composition function should result in representations that are closer to the defined words rather than the other alternatives (excluding synonyms and ambiguously defined words acting as noise). Note that the candidate list includes words with a great deal of diversity in terms of specificity. In order to produce positive and negative similarity instances, we have followed a series of careful steps to avoid variants of the target words, stopwords, multiwords, and so forth, that may introduce noise and deviate the results: (i) Select the set of SINGLE words in the Word2Vec vocabulary (without digits) that do not appear in the MEN dataset and have at least one WordNet synset (80,058 tokens); (ii) Lemmatize the available words according to their grammatical tagging (NOUN, VERB, ADVERB, and ADJECTIVE) (44,642 tokens); (iii) Filter out those lemmas that are stopwords, or have no associated synset in Wordnet, or vector in Word2Vec (43,469 tokens); (iv) Keep 30,000 random words from the filtered lemmas (using a pre-defined seed for reproducibility purposes) (30,000 tokens); (v) Collect the 20,183 different words appearing in MEN; (vi) For each word in MEN, consider the word and its corresponding definition as positive similarity instances (20,183 positive instances). Thus, they are represented by a maximum similarity value of 1; (vii) For each word in MEN, consider their corresponding definition (and we set a maximum similarity among them) and 100 random tokens out of the 30,000 tokens batch collected in step (4) as negative similarity samples (201,830 negative instances). In this second case, they are represented by a similarity value of 0; and (viii) For each similarity function, receiver operator characteristic (ROC) area under the curve (AUC) score is computed taking the 0–1 ground truth labels and the similarity predictions as input.

6.3 Semantic Representation Systems in Comparison

This subsection briefly explains the reasons behind all considered instantiations of the functions composing an ICDS representation system (see Section 5 for the theoretical details and constraints of such systems). These components correspond to the embedding π, composition $\odot$ and the similarity function δ. Table 3 illustrates all possible configuration set-ups that are considered in our experiments.

We consider both static (non-contextual) and contextual embedding functions. The former is instantiated with Word2Vec representations, although other well-known alternatives such as GloVe or FastText (Mikolov et al. 2018) could also apply. In addition, the embedding function in the experiments is also instantiated by pre-trained contextualized representation models based on Transformer Encoder/Decoder architectures: BERT (Devlin et al. 2019) and GPT-2 (Brown et al. 2020). Following the suggestions given by Ethayarajh (2019), we take the second layer in GPT-2. This author suggests that upper layers of contextualizing models produce more context-specific representations.

We also include S-BERT (Reimers and Gurevych 2019) in the experiments. S-BERT is a supervised BERT variant representation approach that provides embeddings at the

Table 3
Configurations considered for the experiments.

Embedding π	**Composition** $\odot$	**Linguistic structures**	**Similarity** δ
GPT-2 (Layer-2)	F_{Avg}, F_{Ind} F_{Joint}, F_{Inf}	Sequential (L2R and R2L)	Cosine Dot product ICM_β Euclidean
	F_{Sum}, Max Pooling	None	
Unsupervised BERT	F_{Avg}, F_{Ind} F_{Joint}, F_{Inf}	Sequential (L2R and R2L)	
	F_{Sum}, Max Pooling, CLS Token	None	
Supervised S-BERT	Global Average		
Word2Vec	F_{Sum}		
	F_{Avg}, F_{Ind} F_{Joint}, F_{Inf}	Sequential, Constituents, Dependency Tree (L2R and R2L)	

sentence level by re-training the model on SNLI and NLI annotated corpora (Bowman et al. 2015), or even on STS training datasets.[6] In terms of our theoretical framework, S-BERT consists of the selected BERT version (BERT, distilBERT, RoBERTa…with re-trained weights) as the embedding function π, followed by a simple composition function $\odot$ (vector global average). S-BERT and similar supervised approaches have reported state-of-the-art results in semantic textual similarity, jointly performing the processes of vector space projection and semantic composition. We are interested in analyzing the relative performance of unsupervised and supervised BERT as well as in determining how task-specific this fine-tuned model is.

Regarding the composition function $\odot$, in our experiments, we consider traditional sum (F_{Sum}) and pairwise average (F_{Avg}) operations, as well as three other configurations of the Generalized Composition Function: F_{Ind}, F_{Joint}, and F_{Inf}. In addition, with regard to the BERT model, we consider max pooling[7] and CLS token extraction[8] as two alternatives to combine the contextualized token representations. All of these composition functions are applied on top of all embedding functions except for S-BERT.

Note that S-BERT simply uses the global average on top of the BERT architecture as a single composition function option.

Concerning the linguistic structures and composition order used in our experiments (see Section 5.4), in the case of BERT and GPT-2, we consider the sequential structure in both L2R and R2L directions. Note that F_{sum}, Max Pooling, Global Average, and CLS token are composition criteria that do not comply with Property 6 (SENSITIVITY TO

6 These datasets were created to identify a certain semantic similarity/relation between pairs of sentences.
7 Max Pooling is a technique commonly used in Computer Vision to reduce the dimension of the image while maintaining the spatial relation of pixels as much as possible. In this NLP operation, we apply it over the dimensions of the token embeddings.
8 CLS token, also known as Classification Token, is a token added at the beginning of each text input during the pre-processing step to be used in classification tasks adopting the role of representing the whole text.

Table 4
Example of the resulting flattened linguistic structures.

Sequential	['Just', 'Robert', 'showed', 'up', 'for', 'the', 'party']
Constituents	[['Just', 'Robert'], ['showed', 'up', ['for', ['the', 'party']]]]
Dependency (L2R)	['showed', ['Robert', 'Just'], 'up', ['party', 'for', 'the']]
Dependency (R2L)	[['Just', 'Robert'], 'up', ['for', 'the', 'party'], 'showed']

STRUCTURE). Therefore, the structure does not affect the composite representation. In the case of static embeddings (Word2Vec), we use sequential, constituency, and dependency parsing. In order to extract the proper linguistic structures from the whole set of texts, we used the Stanford Core NLP API, including both tokenizer and parsers for extracting constituent and dependency trees. These tree representations were later flattened into nested lists following a right-to-left (R2L) or left-to-right (L2R) direction for a more efficient computation of the compositions. A visualization of such flattening output is presented in Table 4.

Finally, the similarity functions included in our experiments are cosine similarity, dot product, ICM_β, and Euclidean similarity. Their definitions, correspondences, and formal properties are described in Section 5.

6.4 Results

For clarity of exposition, the complete table of results will be available in the Github repository.[9] This section shows aggregated data in bar charts that allow us to analyze the results and draw conclusions. The most important empirical results that we have been able to derive from the experiments are described below.

6.4.1 Contextual vs. Static Embedding Functions. According to previous studies (Ethayarajh 2019; Gao et al. 2019; Li et al. 2020), although Contextual Neural Language Models such as BERT and GPT have a high prediction power as Language Models, they do not distribute words isometrically. That is, there is no isometry between the angular position of utterance embeddings and their expected semantic similarity according to humans (ANGULAR ISOMETRY). Consequently, applying these embedding functions may affect the effectiveness of similarity functions, at least in a non-supervised representation oriented experiment such as this. We can observe this effect in our experiments.

Figure 6 shows the Pearson correlation between estimated and real similarities for the datasets STS, SICK, and DEF2DEF. In this experiment, we compare the performance of embedding functions. In every approach, we apply the sequential structure SEQ (selecting the best processing direction, L2R or R2L, in each case), the best composition function (within $\mathtt{F}_{\mathtt{sum}}$, $\mathtt{F}_{\mathtt{avg}}$, $\mathtt{F}_{\mathtt{ind}}$, $\mathtt{F}_{\mathtt{joint}}$, and $\mathtt{F}_{\mathtt{inf}}$), and cosine as similarity functions. In

9 Source code, data used in the experiments, and the complete table of results are publicly available to encourage reproducible research: `https://github.com/alarca94/ICDS`.

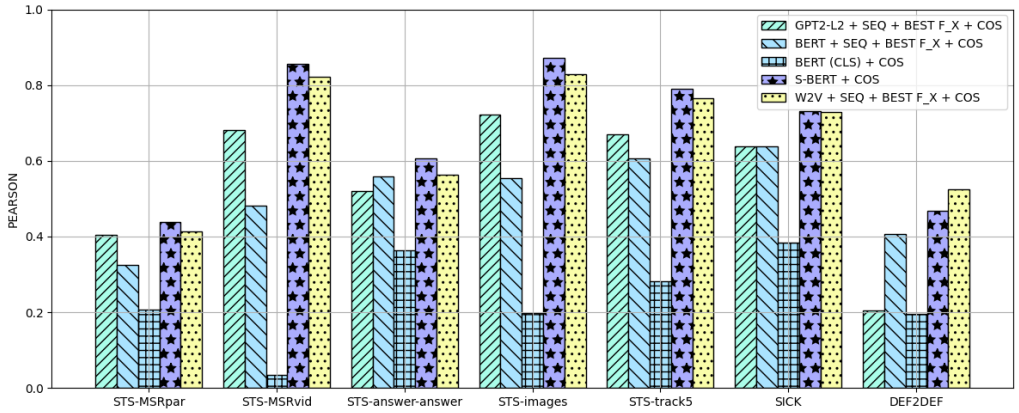

Figure 6
Contextual vs. Static Embedding function performance in terms of Pearson correlation across several semantic similarity prediction datasets using the cosine similarity, a sequential structure, and the best configuration of the remaining components.

Table 5
Contextual vs. Static Embedding function performance in terms of ROC-AUC score at the WORD2DEF dataset using the cosine similarity, a sequential structure, and the best configuration of the remaining components.

		SEQ + COS		
GPT-2	BERT	BERT (CLS)	S-BERT	W2V
0.4713	0.6609	0.5133	0.8841	**0.9589**

the case of BERT-CLS and S-BERT, the linguistic structure and the composition function is given directly by the Neural Language Model. Table 5 shows the result of the same experiment over the WORD2DEF test set and ROC-AUC as an evaluation measure.

As Table 5 and Figure 6 show, unsupervised contextual embedding functions (pre-trained BERT and GPT-2) achieve a lower correlation than static embeddings (Word2Vec) in every dataset. For example, differences of around 50% are observed between BERT and Word2Vec in STS-MSRvid or between GPT-2 and Word2Vec in DEF2DEF. In fact, the results obtained by a two-tailed Steiger's test (Steiger 1980) show a significant statistical difference when comparing Word2Vec to GPT-2 and BERT with a p-value below 2% for all datasets except STS-MSRpar and STS-answer-answer. The correlation values obtained for BERT using the CLS field are even lower and the statistical test shows a significant difference with respect to Word2Vec in all datasets with a p-value of 0%. The differences between contextual and static embedding are more evident in the WORD2DEF dataset where a word is compared to a complete sentence (Table 5).

On the other hand, contextual embeddings perform better when applying a supervised transfer learning from paraphrasing datasets. This is the case of S-BERT, whose correlation value exceeds those obtained with Word2Vec word embeddings. Nevertheless, the statistical test is only able to detect a significant improvement with respect to static embeddings for the STS-images dataset where, similar to other STS subsets and the SICK dataset, sentence pairs are comparable and words overlap. On the contrary,

S-BERT falls behind Word2Vec performance in the DEF2DEF corpus with a p-value of 0. In the case of WORD2DEF, S-BERT achieves an ROC-AUC score of 0.88 while Word2Vec achieves 0.94. Consistent with previous studies (Raffel et al. 2020; Yogatama et al. 2019), this result suggests that transferring learning from supervised tasks can bias the system performance.

In sum, the experiments suggest that *static embedding functions are more effective as text representation models in nonsupervised scenarios*. In our opinion, this is related to the lack of ANGULAR ISOMETRY due to the representation degradation problem in contextual embedding functions.

6.4.2 Composition Functions. In order to analyze the performance of the composition functions, we evaluate static embeddings (Word2Vec) with the best linguistic structure for each dataset and cosine similarity. Figure 7 and Table 6 show the results.

First, we observe that in general, $\mathtt{F}_{\mathtt{avg}}$ performs worse than other functions. In particular, Steiger's method on Pearson correlations indicates that $\mathtt{F}_{\mathtt{avg}}$ is statistically outperformed by $\mathtt{F}_{\mathtt{ind}}$, $\mathtt{F}_{\mathtt{inf}}$, and $\mathtt{F}_{\mathtt{joint}}$ with a p-value below 3% on 5 out of 7 datasets with the exception of STS-MSRvid and STS-answer-answer. This finding suggests that COMPOSITION NEUTRAL ELEMENT is a relevant property (see Table 2). Note that $\mathtt{F}_{\mathtt{avg}}$ has the counter-intuitive effect of reducing the magnitude (Information Content) of the embedding as we add information by compositing it with other shorter vectors.

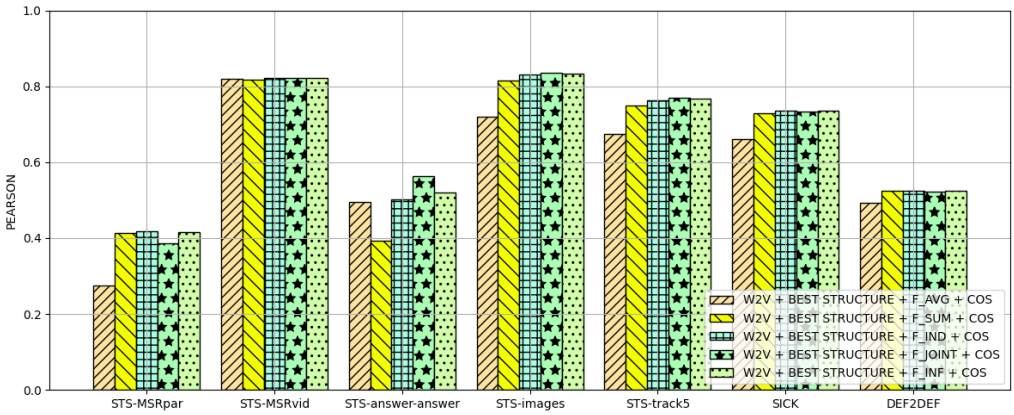

Figure 7
Composition function performance in terms of Pearson correlation across several semantic similarity prediction datasets using static embeddings, cosine similarity, and the best configuration of the remaining components.

Table 6
Composition function performance in terms of ROC-AUC score at the WORD2DEF dataset using static embeddings, cosine similarity, and the best configuration of the remaining components.

W2V + BEST STRUCTURE + COS				
$\mathtt{F}_{\mathtt{AVG}}$	$\mathtt{F}_{\mathtt{SUM}}$	$\mathtt{F}_{\mathtt{IND}}$	$\mathtt{F}_{\mathtt{JOINT}}$	$\mathtt{F}_{\mathtt{INF}}$
0.9511	0.9589	0.9595	**0.9597**	0.9596

In the datasets DEF2DEF and WORD2DEF, without word overlap between sentences, $\mathtt{F}_{\mathtt{sum}}$ obtains similar results to the non-associative functions $\mathtt{F}_{\mathtt{ind}}$, $\mathtt{F}_{\mathtt{joint}}$, or $\mathtt{F}_{\mathtt{inf}}$. These results suggest that the SENSITIVITY TO TEXT STRUCTURE in those datasets is not crucial when composing embeddings. However, in the case of sentence-level datasets with a high overlap (SICK and 3 out of 6 STS datasets), the structure-sensitive composite embedding functions, $\mathtt{F}_{\mathtt{ind}}$, $\mathtt{F}_{\mathtt{joint}}$, and $\mathtt{F}_{\mathtt{inf}}$ outperform $\mathtt{F}_{\mathtt{sum}}$ with a p-value close to 0%. In sum, as can be guessed, *satisfying* SENSITIVITY TO TEXT STRUCTURE *in the composition function is especially relevant when comparing sentences with high word overlap.*

We do not find a consistent and substantial difference across datasets between $\mathtt{F}_{\mathtt{inf}}$ and the parameterizations at the boundaries of the theoretical area in which COMPOSITION NORM LOWERBOUND and COMPOSITION NORM MONOTONICITY are satisfied (see Table 2). However, it should be noted that $\mathtt{F}_{\mathtt{joint}}$ and $\mathtt{F}_{\mathtt{ind}}$ are just at the boundaries (see Section 5.2).

6.4.3 Similarity Functions. In the previous analysis, the similarity function was fixed to be the cosine similarity in order to consistently explore the composition function behavior. Now, we explore the alternative similarity functions. Table 7 and Figure 8 contain the average performance across datasets for S-BERT and Word2Vec with different composition functions. In the case of static embedding (Word2Vec), we selected the best linguistic structure per dataset. Note that in order to average Pearson correlations, scores are transformed with Fisher Z-transformation, averaged and transformed back to the original space.

First, the results provide evidence for the importance of satisfying the formal constraints ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY and EQUIDISTANT EMBEDDING SIMILARITY MONOTONICITY. Note that ICM falls within the limits of the theoretical range corresponding to the dot product (ICM with $\beta = 1$) and Euclidean distance (ICM with $\beta = 2$). In all static embedding cases, ICM outperforms the dot product and Euclidean distance. To be more precise, when ICM is paired with $\mathtt{F}_{\mathtt{inf}}$ so that all considered properties are satisfied, a statistical significant difference is achieved in all 7 datasets with respect to dot product and 4 datasets when comparing it against Euclidean distance.

However, the cosine distance obtains greater results than ICM with the theoretically estimated parameter with statistical significance in 3 to 4 datasets out of the 7 datasets included in Figure 8, depending on the composition function used to compute final

Table 7
Similarity function performance in terms of ROC-AUC score at the WORD2DEF dataset per composition function on static embeddings after selecting their best configuration per dataset. S-BERT is added for comparison purposes (contextual embeddings and state-of-the-art performance).

	COS	DOT	ICM	EUC
S-BERT	0.8840	0.8819	**0.9447**	0.8837
W2V + $\mathtt{F}_{\mathtt{AVG}}$	0.9511	0.9340	**0.9550**	0.7643
W2V + $\mathtt{F}_{\mathtt{IND}}$	**0.9595**	0.9179	0.9586	0.6722
W2V + $\mathtt{F}_{\mathtt{JOINT}}$	0.9597	0.9320	**0.9617**	0.7465
W2V + $\mathtt{F}_{\mathtt{INF}}$	0.9596	0.9253	**0.9605**	0.7070
W2V + $\mathtt{F}_{\mathtt{SUM}}$	**0.9589**	0.8926	0.9519	0.5981

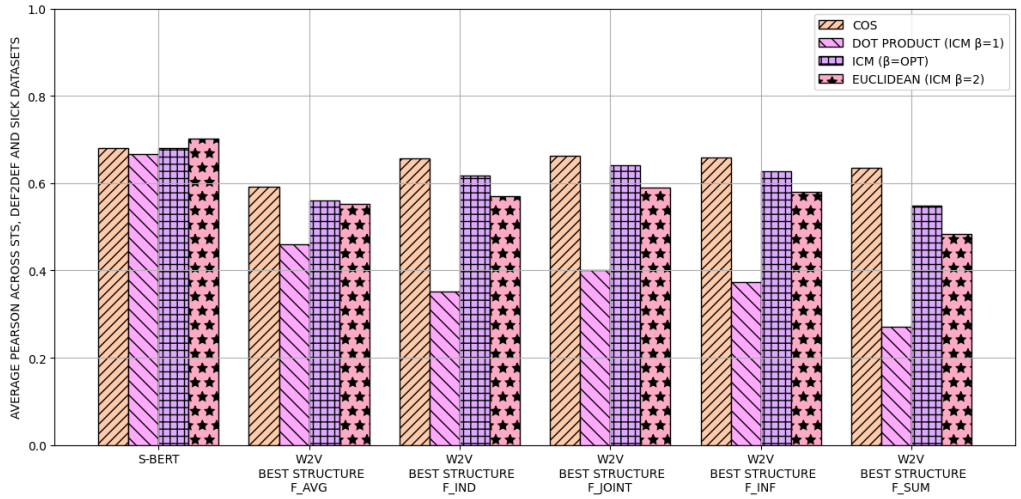

Figure 8
Similarity function average performance in terms of Pearson correlation over all presented semantic similarity datasets per composition function on static embeddings after selecting their best configuration per dataset. S-BERT is added for comparison purposes (contextual embeddings and state-of-the-art performance).

embeddings. Based on our embedding space interpretation (see Section 3.1), this result suggests that the Information Content or specificity of the compared texts are not crucial in our semantic similarity datasets. In other words, the annotations in at least these textual similarity datasets are more focused on angular distance (pragmatic similarity) than vector magnitudes (literal similarity which takes into account the amount of information provided by texts).

The difference between both similarity functions is really subtle in the case of WORD2DEF lying on both directions depending on which composition function is presented, and thus no clear conclusions can be drawn. One possible reason why both similarities are on par with each other in this scenario when previous datasets presented significant differences could be the increasing importance of ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY when specificity between each pair of texts differs in larger quantities.

6.4.4 Linguistic Structures. One of the premises of this research work is the fact that following a structure in the composition stage of an ICDS representation system should benefit performance with respect to traditional orderless baselines such as the sum of vectors. This statement was proven to be true in Section 6.4.2 where the sum operation was outperformed by all other unsupervised composition functions that followed a certain structure. The next experiment aims to investigate whether there are significant performance changes when alternating from one linguistic structure to another among all of the options presented in Section 5.4. The corresponding results for each structure across datasets are shown in Figure 9, fixing the model configuration to Word2Vec as the representation system, $\mathtt{F}_{\mathtt{inf}}$ as the composition function, and cosine as the similarity function. The ROC-AUC results for the dataset WORD2DEF are 0.95847, 0.95967, and 0.95941 for the sequential structure, syntactic constituents, and dependency trees, respectively.

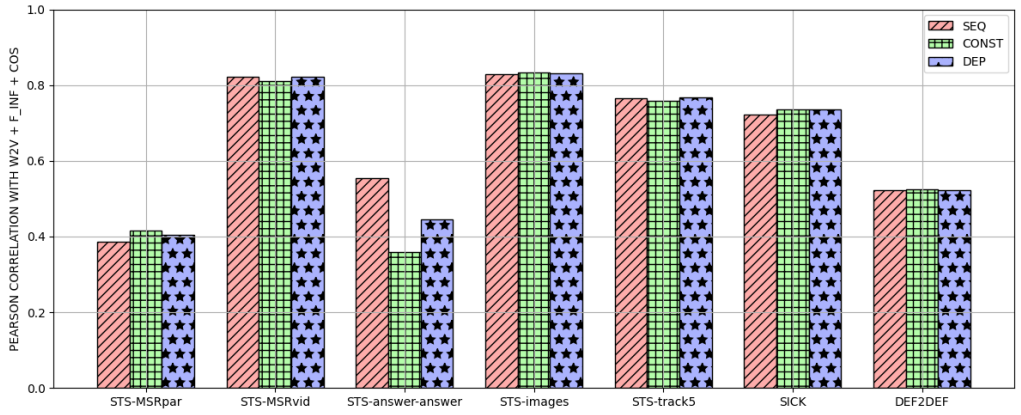

Figure 9
Similarity prediction power of our main proposed baseline (Word2Vec as the representation system, F_{Inf} as the composition function and cosine as the similarity function) in terms of Pearson correlation across all presented semantic similarity datasets with respect to the selected linguistic structure to follow during composition.

Unfortunately, no substantial difference can be found among the tested linguistic structures. The effect of each linguistic structure depends strongly on the dataset and task. For instance, both DEP and SEQ achieve a significant statistical difference with respect to CONST at STS-MSRvid with a p-value below 5%; but SEQ is outperformed significantly by DEP and CONST at SICK with a p-value of 0% and both DEP and CONST are significantly outperformed by SEQ, with a p-value below 5% at STS-answer-answer. No other significant differences can be extracted from the statistical analysis for the other 4 datasets. Nonetheless, the importance and benefits of following a structure instead of associative composition functions has been illustrated previously.

7. Conclusions

In this article, we have formalized the concept of an Information Theory–based Compositional Distributional Semantics as a tuple of three functions, namely, embedding, composition, and similarity. The main contribution is to have established a bridge between representation space and Information Theory, not only for embeddings but also for composition and similarity functions. The proposed framework includes a set of nine formal constraints based on Information Content boundaries and monotonicity in embedding, composition, and similarity functions.

Regarding embedding function, our experiments show that adding embeddings from internal layers of contextual neural networks such as GPT-2 or BERT does not outperform the use of static embeddings when performing a textual semantic composition process, at least in terms of the semantic similarity prediction. These results are consistent with the Representation Degradation Problem identified in the literature (Ethayarajh 2019; Gao et al. 2019; Li et al. 2020). It is true that the S-BERT model based on supervised transfer learning manages to increase the effectiveness in certain datasets. However, its effectiveness decreases in problems that are not similar to the tasks on which it has been trained, as in the case of word-definition comparison.

Our formal analysis shows that popular, well-known baseline composition and similarity functions do not satisfy all the constraints simultaneously. In order to overcome this formal limitation, we have defined two parameterizable approaches that generalize standard composition (sum) and similarity (Euclidean distance or dot product) functions. Our theoretical study shows that these functions satisfy all constraints within certain parameter ranges. Our empirical study suggests that the generalized composition function within the theoretical parameter range outperforms standard functions such as the sum or the global average.

The same occurs in the case of the similarity functions, where the Euclidean distance or the scalar product fall outside the range. However, our results also show that the cosine distance, while not sensitive to the vector norm (Property 8, ORTHOGONAL EMBEDDING SIMILARITY MONOTONICITY), is a robust similarity estimate. According to our embedding space interpretation, this means that the similarity annotations in our datasets are more closely related with *the pragmatic context* (i.e., vector direction) than with *literal utterance similarity*. That is, it does not take into account how much information the message provides in relation to its meaning in context (i.e., vector norm).

Finally, according to our experiments, the election of linguistic structures does not affect the performance substantially. Sequential text processing achieves results similar to those of more complex structures. However, the results also show that considering structure (sequential, constituents, or dependencies) outperforms those approaches that do not consider word order (i.e. sum, or global average of embeddings). That is, we have not been able to contribute new results to the discussion between the sequential or hierarchical nature of language, although we have been able to confirm that considering structures is more effective than representing texts as a bag of words. It should be noted that this effect becomes tangible in the case of similarity between sentences with some overlap of words (STS and SICK corpora), whereas in our definition vs. definition or definition vs. word comparison, we have not managed to observe this effect.

In view of our analysis, our recommendation for unsupervised text representation is to combine static (non-contextual) embeddings following a sequential structure using the proposed composition function $\mathtt{F}_{\mathtt{inf}}$, and using ICM with either the estimated β value or cosine as a similarity function. Note that this recommendation applies only to non-supervised scenarios. The literature has demonstrated repeatedly the predictive power of contextual models in supervised tasks.

As future work we can identify several research lines. Firstly, to address the problem of the degeneration of representation in contextual approaches. This direction has already been explored in multiple studies (Ethayarajh 2019; Gao et al. 2019; Li et al. 2020). Also, in line with previous work, we see the need to design semantic similarity datasets over more heterogeneous cases and scenarios when comparing general purpose representation models. A third line of work is the development of datasets that capture the effect of specificity and the amount of information on the estimation of semantic similarity between textual representations, as has been observed in experiments within cognitive sciences (Tversky 1977; Amigó et al. 2020). A fourth and perhaps the most ambitious line of research is to close the circle of *Neuro-Symbolic Artificial Intelligence* in the context of distributional semantics. In this article we have addressed the representation phase in which symbolic information (text) is translated into a continuous space. The pending step is *extraction*, in which the representation in this continuous space is translated into a symbolic domain. There are several authors who have focused on this challenge (Sarker et al. 2021).

Appendix A. Proofs

A.1 Proof for Theorem 1

The generalized composition function is defined as

$$\mathtt{F}_{\lambda,\mu}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|} \cdot \sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}$$

If $\|\vec{v}_2\| = 0$ then $\mathtt{F}_{\lambda,\mu}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1}{\|\vec{v}_1\|} \cdot \sqrt{\lambda(\|\vec{v}_1\|^2)} = \frac{\vec{v}_1}{\|\vec{v}_1\|} \cdot \sqrt{\lambda}\|\vec{v}_1\|$, which is equal to $\vec{v}_1$ whenever $\lambda = 1$, satisfying Composition Neutral Element. Also, if $\|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\|$ then

$$\|\mathtt{F}_{\lambda,\mu}(\vec{v}_1, \vec{v}_2)\| = \sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_1\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle} = \|\vec{v}_1\| \cdot \sqrt{2\lambda - \mu \cdot \cos(\vec{v}_1, \vec{v}_2)}$$

Therefore, the composite embedding magnitude is inversely monotonic regarding the cosine whenever $\mu > 0$, satisfying Composition Norm Monotonicity. So, at this point, we can state the constraints $\lambda = 1$ and $\mu > 0$.

Let us analyze the Composition Norm Lowerbound property. The square of the composite embedding magnitude (information content) is $\|\mathtt{F}_{\lambda,\mu}(\vec{v}_1, \vec{v}_2)\|^2 = \lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle = \lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu(\|\vec{v}_1\| \cdot \|\vec{v}_2\| \cdot \cos(\vec{v}_1, \vec{v}_2))$.

Given that $\lambda = 1$, $\mu > 0$, and $\cos(\vec{v}_1, \vec{v}_2) \leq 1$, we can assert that $\|\mathtt{F}_{\lambda,\mu}(\vec{v}_1, \vec{v}_2)\|^2 \geq \|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \mu(\|\vec{v}_1\| \cdot \|\vec{v}_2\|)$. If the magnitude of the composite embedding equals the magnitude of the maximum component then

$$\begin{aligned}
&\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \mu\|\vec{v}_1\|\|\vec{v}_2\| = \max(\|\vec{v}_1\|, \|\vec{v}_2\|)^2 \\
&\iff \max(\|\vec{v}_1\|, \|\vec{v}_2\|)^2 + \min(\|\vec{v}_1\|, \|\vec{v}_2\|)^2 - \mu\|\vec{v}_1\|\|\vec{v}_2\| = \mathit{max}(\|\vec{v}_1\|, \|\vec{v}_2\|)^2 \\
&\iff \mu = \frac{\min(\|\vec{v}_1\|, \|\vec{v}_2\|)^2}{\|\vec{v}_1\|\|\vec{v}_2\|} = \frac{\min(\|\vec{v}_1\|, \|\vec{v}_2\|)}{\max(\|\vec{v}_1\|, \|\vec{v}_2\|)}
\end{aligned}$$

Therefore $\|\mathtt{F}_{\lambda,\mu}(\vec{v}_1, \vec{v}_2)\| \geq \max(\|\vec{v}_1\|, \|\vec{v}_2\|)$ and Vector Magnitude Composition Monotonicity is satisfied whenever $\mu \in \left(0, \frac{\min(\|\vec{v}_1\|, \|\vec{v}_2\|)^2}{\|\vec{v}_1\|\|\vec{v}_2\|}\right]$.

Finally, let us analyze the Sensitivity to Structure property. Because $\|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\| > 0$ and $\cos(\vec{v}_1, \vec{v}_2) = \cos(\vec{v}_1, \vec{v}_3) = \cos(\vec{v}_2, \vec{v}_3) > 0$, then the proposition $(\vec{v}_1 \odot \vec{v}_2) \odot \vec{v}_3 \neq \vec{v}_1 \odot (\vec{v}_2 \odot \vec{v}_3)$ fits at least if their unit vectors are different. Therefore, we only need to check that

$$\frac{\vec{v}_3 + \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|}}{\left\|\vec{v}_3 + \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|}\right\|} \neq \frac{\vec{v}_1 + \frac{\vec{v}_2 + \vec{v}_3}{\|\vec{v}_2 + \vec{v}_3\|}}{\left\|\vec{v}_1 + \frac{\vec{v}_2 + \vec{v}_3}{\|\vec{v}_2 + \vec{v}_3\|}\right\|}$$

Given that the three vectors share the same norm and relative angular distance (cosine), we can assert that $\|\vec{v}_2 + \vec{v}_3\| = \|\vec{v}_1 + \vec{v}_2\|$ and

$$\left\|\vec{v}_3 + \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|}\right\| = \left\|\vec{v}_1 + \frac{\vec{v}_2 + \vec{v}_3}{\|\vec{v}_2 + \vec{v}_3\|}\right\|$$

Therefore, we only need to prove that $\vec{v}_3 + \frac{\vec{v}_1+\vec{v}_2}{c} \neq \vec{v}_1 + \frac{\vec{v}_2+\vec{v}_3}{c}$, for which it is a sufficient condition that it is fulfilled at least for one of the vector components. That is to say, $\exists i \left(v_{3_i} + \frac{v_{1_i}+v_{2_i}}{c} \neq v_{1_i} + \frac{v_{2_i}+v_{3_i}}{c}\right)$. This is equivalent to $\exists i\,(c \cdot v_{3_i} + v_{1_i} + v_{2_i} \neq c \cdot v_{1_i} + v_{2_i} + v_{3_i})$, which is true whenever $(v_{3_i} \neq v_{1_i})$, since the property states as a condition that the cosine between vectors is greater than zero as well as their norms. Therefore, we can affirm that $(v_{3_i} \neq v_{1_i})$ for at least one component of the vector, so it is demonstrated that the composition function satisfies the SENSITIVITY TO STRUCTURE property.

A.2 Formal Proof for Theorem 2

The embedding based ICM metric is defined as

$$ICM^V_\beta = (1-\beta)(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) + \beta\|\vec{v}_1\|\|\vec{v}_2\| \cos(\vec{v}_1, \vec{v}_2)$$

In order to satisfy ANGULAR ISOMETRICITY, the expression must be monotonic regarding $\cos(\vec{v}_1, \vec{v}_2)$ and, therefore, β must be strictly greater than zero. In order to satisfy COMPOSITION NORM MONOTONICITY, ICM^V_β must decrease with $\|\vec{v}_1\|$ and $\|\vec{v}_2\|$ when $\cos(\vec{v}_1, \vec{v}_2) = 0$. Therefore $\beta > 1$.

Let us consider SENSITIVITY TO STRUCTURE. Let $\vec{v} = (x_1, .., x_n)$ and $\vec{\epsilon} = (\epsilon_1, .., \epsilon_n)$. Then

$$ICM^V_\beta(\vec{v}, \vec{v}+\vec{\epsilon}) = \|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \beta\left(\|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \langle\vec{v}, \vec{\epsilon}\rangle\right) =$$

$$= \|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \beta\sum_i (x_i^2 + (x_i+\epsilon_i)^2 - x_i(x_i+\epsilon_i)) = \|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \beta\sum_i (x_i^2 + x_i^2 + \epsilon_i^2 + 2x_i\epsilon_i - x_i^2 - x_i\epsilon_i)) =$$

$$= \|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \beta\sum_i (x_i^2 + \epsilon_i^2 + x_i\epsilon_i)) = \|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \beta\sum_i \tfrac{1}{2}(x_i^2 + \epsilon_i^2) + \sum_i \tfrac{1}{2}(x_i^2 + \epsilon_i^2 + 2x_i\epsilon_i)) =$$

$$= \|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \beta\left(\tfrac{1}{2}\|\vec{v}\|^2 + \tfrac{1}{2}\|\vec{\epsilon}\|^2 + \tfrac{1}{2}\|\vec{v}+\vec{\epsilon}\|^2\right)$$

We remove the constant $\frac{1}{2}\|\vec{\epsilon}\|^2$. We obtain $(\|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2 - \beta\left(\frac{1}{2}\|\vec{v}\|^2 + \frac{1}{2}\|\vec{v}+\vec{\epsilon}\|^2\right)) = (1-\frac{\beta}{2})(\|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2)$. Assuming that $\|\vec{v}\| >> \|\vec{\epsilon}\|$ then $\|\vec{v}+\vec{\epsilon}\|^2$ is monotonic regarding $\|\vec{v}\|$. Therefore, the expression $(1-\frac{\beta}{2})(\|\vec{v}\|^2 + \|\vec{v}+\vec{\epsilon}\|^2)$ is monotonic regarding $\|\vec{v}\|$ whenever $\beta < 2$, satisfying SENSITIVITY TO STRUCTURE.

Acknowledgments

The authors would like to thank anonymous reviewers for their valuable suggestions to improve the quality of the article. This research was partially supported by the Spanish Ministry of Science and Innovation Project FairTransNLP (PID2021-124361OB-C32); by MCI/AEI/FEDER, UE DOTT-HEALTH Project under grant PID2019-106942RB-C32; and LyrAIcs project through the European Research Council (ERC), under the Research and Innovation Program Horizon2020 under grant agreement number 964009. The work has been carried out in the framework of the following projects: MISMIS project (PGC2018-096212-B), funded by Ministerio de Ciencia, Innovación y Universidades (Spain) and CLiC SGR (2027SGR341), funded by AGAUR (Generalitat de Catalunya).

References

Agirre, Eneko and German Rigau. 1996. Word sense disambiguation using conceptual density. In *COLING 1996 Vol. 1: The 16th International Conference on Computational Linguistics*, pages 16–22. https://doi.org/10.3115/992628.992635

Almuhareb, Abdulrahman. 2006. *Attributes in lexical acquisition*. Ph.D. thesis, University of Essex, Colchester, UK.

Amigó, Enrique, Fernando Giner, Julio Gonzalo, and M. Verdejo. 2020. On the foundations of similarity in information

access. *Information Retrieval Journal*, 23:216–254. https://doi.org/10.1007/s10791-020-09375-z

Amigó, Enrique and Julio Gonzalo. 2022. An empirical study on similarity functions: Parameter estimation for the information contrast model. OSF Preprints. https://doi.org/10.31219/osf.io/3b27t

Andreas, Jacob, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. In *Proceedings of the 51st ACL (Vol. 2: Short Papers)*, pages 47–52.

Arora, Sanjeev, 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. https://doi.org/10.1162/tacl_a_00106

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 238–247. https://doi.org/10.3115/v1/P14-1023

Baroni, Marco, Stefan Evert, and Alessandro Lenci. 2008. Esslli workshop on distributional lexical semantics bridging the gap between semantic theory and computational simulations. http://wordspace.collocations.de/doku.php/data:esslli2008:start.

Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721. https://doi.org/10.1162/coli_a_00016

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Blacoe, William and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556.

Blei, David M., Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Boleda, Gemma and Katrin Erk. 2015. Distributional semantic features as semantic primitives—or not. In *AAAI Spring Symposium Series*, pages 2–5.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on EMNLP*, pages 632–642. https://doi.org/10.18653/v1/D15-1075

Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. 2014a. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47. https://doi.org/10.1613/jair.4135

Bruni, Elia, Nam Khanh Tran, and Marco Baroni. 2014b. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47. https://doi.org/10.1613/jair.4135

Cai, Xingyu, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Camacho-Collados, José and Roberto Navigli. 2016. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for*

NLP, pages 43–50. https://doi.org/10.18653/v1/W16-2508

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. https://doi.org/10.18653/v1/S17-2001

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175. https://doi.org/10.18653/v1/D18-2029

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Clark, Stephen and Stephen G. Pulman. 2007. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394.

Collobert, Ronan and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Machine Learning, Proceedings of the ICML*, pages 160–167. https://doi.org/10.1145/1390156.1390177

Cummins, Robert. 1996. Systematicity. *The Journal of Philosophy*, *93*:591–614. https://doi.org/10.2307/2941118

Czarnowska, Paula, Guy Emerson, and Ann Copestake. 2019. Words are vectors, dependencies are matrices: Learning word embeddings from dependency graphs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 91–102. https://doi.org/10.18653/v1/W19-0408

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Demeter, David, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2191–2197. https://doi.org/10.18653/v1/2020.acl-main.198

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1*, pages 4171–4186.

Erk, Katrin. 2009. Supporting inferences in semantic space: Representing words as regions. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 '09, pages 104–115. https://doi.org/10.3115/1693756.1693769

Ethayarajh, Kawin. 2018. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 91–100. https://doi.org/10.18653/v1/W18-3012

Ethayarajh, Kawin. 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. https://doi.org/10.18653/v1/D19-1006

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books. https://doi.org/10.7551/mitpress/7287.001.0001

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 116–131. https://doi.org/10.1145/371920.372094

Firth, J. R. 1957. *Papers in Linguistics, 1934–1951*. Oxford University Press, London.

Frank, Stefan, Rens Bod, and Morten Christiansen. 2012. How hierarchical is language use? *Proceedings. Biological Sciences / The Royal Society*, 279:4522–4531.

Gao, Jun, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in

training natural language generation models. In *Proceedings of ICLR*.

Gerz, Daniela, Ivan Vulic, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *EMNLP*, pages 2173–2182. https://doi.org/10.18653/v1/D16-1235

Gladkova, Anna and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42. https://doi.org/10.18653/v1/W16-2507

Goodwin, Emily, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 1958–1969. https://doi.org/10.18653/v1/2020.acl-main.177

Grefenstette, Edward and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the EMNLP '11*, pages 1394–1404.

Harris, Zellig. 1954. Distributional structure. *Word*, 10(2–3):146–162. https://doi.org/10.1080/00437956.1954.11659520

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. https://doi.org/10.1162/COLI_a_00237

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalize? *JAIR*, 67:757–795. https://doi.org/10.1613/jair.1.11674

Johnson, Kent. 2004. On the systematicity of language and thought. *Journal of Philosophy*, 101(3):111–139. https://doi.org/10.5840/jphil2004101321

Kartsaklis, Dimitri, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In *COLING (Posters)*, pages 549–558.

Kenter, Tom, Alexey Borisov, and Maarten de Rijke. 2016. Siamese CBOW: Optimizing word embeddings for sentence representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 941–951. https://doi.org/10.18653/v1/P16-1089

Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28*, pages 3294–3302.

Landauer, Thomas K. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240. https://doi.org/10.1037/0033-295X.104.2.211

Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Levy, Omer and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems 27*, pages 2177–2185.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461. https://doi.org/10.18653/v1/2020.acl-main.703

Li, Bohan, 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. https://doi.org/10.18653/v1/2020.emnlp-main.733

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the LREC'14*, pages 216–223.

Maruyama, Yoshihiro. 2019. Compositionality and contextuality: The symbolic and statistical theories of meaning. In *Modeling and Using Context - 11th International and Interdisciplinary Conference*, pages 161–174. https://doi.org/10.1007/978-3-030-34974-5_14

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. *11th LREC 2018*, pages 52–55.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings*, pages 240–243. https://doi.org/10.3115/1075812.1075866

Mimno, David and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on EMNLP*, pages 2873–2878. https://doi.org/10.18653/v1/D17-1308

Mitchell, J. and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244.

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x, PubMed: 21564253

Mu, Jiaqi and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69. https://doi.org/10.1145/1459352.1459355

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199. https://doi.org/10.1162/coli.2007.33.2.161

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 EMNLP*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perone, Christian S., Roberto Silveira, and Thomas S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *CoRR*, abs/1806.06259.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *2018 NAACL: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *58th Annual Meeting of the ACL*, pages 4609–4622. https://doi.org/10.18653/v1/2020.acl-main.420

Poerner, Nina, Ulli Waltinger, and Hinrich Schütze. 2020. Sentence meta-embeddings for unsupervised semantic textual similarity. In *Proceedings of the 58th Meeting of the ACL*, pages 7027–7034. https://doi.org/10.18653/v1/2020.acl-main.628

Polajnar, Tamara, Laura Rimell, and Stephen Clark. 2014. Evaluation of simple distributional compositional operations on longer texts. In *Proceedings of the (LREC'14)*, pages 4440–4443.

Qiu, Yuanyuan, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 209–221. https://doi.org/10.1007/978-3-030-01716-3_18

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu,

et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 EMNLP-IJCNLP*, pages 3982–3992. https://doi.org/10.18653/v1/D19-1410

Rimell, Laura, Jean Maillard, Tamara Polajnar, and Stephen Clark. 2016. RELPRON: A relative clause evaluation data set for compositional distributional semantics. In *Computational Linguistics*, 42(4):661–701. https://doi.org/10.1162/COLI_a_00263

Robertson, Stephen. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520. https://doi.org/10.1108/00220410410560582

Rogers, Anna, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148. https://doi.org/10.18653/v1/S17-1017

Salton, Gerard and Michael E. Lesk. 1965. The SMART automatic document retrieval systems—an illustration. *Communications of the ACM*, 8(6):391–398. https://doi.org/10.1145/364955.364990

Sarker, Md Kamruzzaman, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. 2021. Neuro-symbolic artificial intelligence: Current trends. *arXiv preprint arXiv:2105.05330*. https://doi.org/10.3233/AIC-210084

Schnabel, Tobias, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307. https://doi.org/10.18653/v1/D15-1036

Seco, Nuno, Tony Veale, and Jer Hayes. 2004. An intrinsic information content metric for semantic similarity in Wordnet. In *ECAI'04: Proceedings of the 16th European Conference on Artificial Intelligence*, volume 16, pages 1089–1090.

Sekine, Satoshi and Ralph Grishman. 1995. A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the 4th International Workshop on Parsing Technologies*, pages 216–223.

Smolensky, Paul, Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, and Li Deng. 2016. Basic reasoning with tensor product representations. *arXiv preprint arXiv:1601.02745*.

Socher, Richard, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the EMNLP-CoNLL '12*, pages 1201–1211.

Steiger, James H. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87:245–251. https://doi.org/10.1037/0033-2909.87.2.245

Talmor, Alon, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics—On what language model pre-training captures. *Transactions of the ACL*, 8:743–758. https://doi.org/10.1162/tacl_a_00342

Tsvetkov, Yulia, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054. https://doi.org/10.18653/v1/D15-1243

Turney, Peter D. 2007. Empirical evaluation of four tensor decomposition algorithms. *CoRR*, abs/0711.2023. 0711.2023.

Tversky, A. 1977. Features of similarity. *Psychological Review*, 84:327–352. https://doi.org/10.1037/0033-295X.84.4.327

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19. https://doi.org/10.1017/ATSIP.2019.12

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358. https://doi.org/10.1162/tacl_a_00143

Wieting, John and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In

Proceedings of the 56th ACL (Vol. 1: Long Papers), pages 451–462. https://doi.org/10.18653/v1/P18-1042

Wilks, Yorick. 1968. On-line semantic analysis of English texts. *Mechanical Translation and Computational Linguistics*, 11(2):59–72.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Wu, John, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity analysis of contextual word representation models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4638–4655. https://doi.org/10.18653/v1/2020.acl-main.422

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, page 32.

Yogatama, Dani, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *CoRR*, abs/1901.11373.

Zanzotto, Fabio Massimo, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the COLING 2010*, pages 1263–1271.

Zhai, ChengXiang. 2008. Statistical language models for information retrieval a critical review. *Foundations and Trends in Information Retrieval*, 2(3):137–213. https://doi.org/10.1561/1500000008

Zhang, Jiannan. 2014. *A Generic Evaluation of a Categorical Compositional-distributional Model of Meaning*. Ph.D. thesis, University of Oxford.

Zhang, Yan, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on EMNLP*, pages 1601–1610. https://doi.org/10.18653/v1/2020.emnlp-main.124

Zhelezniak, Vitalii, Aleksandar Savkov, and Nils Hammerla. 2020. Estimating mutual information between dense word embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8361–8371. https://doi.org/10.18653/v1/2020.acl-main.741