On the Correspondence between the Squared Norm and Information Content in Text Embeddings

Enrique Amigó, Adrián Ghajari,
Alejandro Benito-Santos and **Diego de la Fuente Rodríguez**
Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

Abstract

Previous work has reported both empirical and theoretical evidence, for specific training models, of the correspondence between the squared norm of an embedding and the information content of the text it represents. In this paper, we investigate the relationship at the theoretical and empirical levels, focusing on the mechanisms and composition functions used to combine token embeddings. i) We formally derive two sufficient theoretical conditions for this correspondence to hold in embedding models. ii) We empirically examine the correspondence and the validity of these conditions at the word level for both static and contextual embeddings and different subword token composition mechanisms. iii) Building on Shannon's Constant Entropy Rate (CER) principle, we explore whether embedding mechanisms exhibit a linearly monotonic increase in information content as text length increases. Our formal analysis and experiments reveal that: i) At the word embedding level, models satisfy the sufficient conditions and show a strong correspondence when certain subword composition functions are applied. ii) Only scaled embedding averages proposed in this paper and certain information-theoretic composition functions preserve the correspondence. Some non-compositional representations—such as the CLS token in BERT or the EOS token in LLaMA—tend to converge toward a fixed point. The CLS token in ModernBERT, however, exhibits behavior that aligns more closely with the CER hypothesis.

1 Introduction

Embeddings remain a cornerstone of modern natural language processing, underpinning a wide range of tasks such as information retrieval, semantic textual similarity, and knowledge graph reasoning. Their utility, however, depends not only on the representations themselves but also on the operators that act upon them, which enable measuring similarity, performing composition, and supporting more complex forms of inference. An equally important aspect is measuring the amount of information embedded in these representations, as it plays a central role in defining and guiding such operators.

Previous studies have provided empirical evidence of a relationship between the squared norm of an embedding and the information content (IC) or self-information ($I(w) = -log(P(w))$) of the corresponding text ($I(w) \propto \|\vec{w}\|^2$). For instance, Levy and Goldberg (2014) analytically demonstrated this correspondence at the word level in the context of the Skip-Gram with Negative Sampling (SGNS) objective. Oyama et al. (2023) further derives a relationship between the squared norm and the Kullback-Leibler (KL) divergence between a word's context distribution and the unigram distribution, extending the analysis to contextual word embeddings (see references in Section 2.1).

In this paper, we investigate the correspondence between IC and the squared norm of embeddings at the text level, focusing on how embedding composition and text-level representations preserve this relationship. To this end, we first identify sufficient (though not necessary) theoretical conditions under which the correspondence holds, regardless of the embedding model or the mechanism used to represent full texts (e.g., CLS token, EOS token, or composition functions). Specifically, the sufficient theoretical conditions identified that support this correspondence are: i) a monotonic relationship between the norm and the amount of information, and ii) that each component of the embedding contributes independently to the estimation of the information content.

We then empirically evaluate these conditions at the word level for both static and contextual embedding models. Our results show that, when the appropriate subword composition function is applied, the models satisfy the sufficient theoreti-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13631–13643
November 4-9, 2025 ©2025 Association for Computational Linguistics

cal conditions and exhibit a strong correspondence between IC and squared norm ($I(w) \propto \|\vec{w}\|^2$).

To analyze the correspondence at the word sequence level, we build on Shannon's Constant Entropy Rate (CER) hypothesis. CER states that when processing a sequence of words (e.g., a sentence), the average entropy per unit (word, symbol, etc.) remains constant as the sequence grows longer. This is equivalent to say that the IC or self-information of the full text increases linearly with its length, at least beyond a certain length (Shannon, 1951). The phenomenon has been corroborated by numerous studies across both psycholinguistics and computational linguistics (see references in Section 2.2).

In this paper, we formally analyze the extent to which word embedding composition functions (e.g., sum, average) preserve the linear monotonicity of the squared norm as words are incrementally added to a text. In parallel, we empirically examine the behavior of various text embedding strategies. Specifically, we compare single-token representations—such as the CLS token in BERT or the last token in LLaMA—with composition-based methods applied to both static and contextual embeddings (see Sections 2.4 and 2.3 for further details).

At a theoretical level, we conclude in Section 4 that neither embedding summation ($\sum \vec{v}_i$) nor averaging ($\frac{\sum \vec{v}_i}{n}$) preserves the linear growth of squared norm (information content). However, normalizing the sum by the square root of the number of tokens $\left(\frac{\sum \vec{v}_i}{\sqrt{n}}\right)$, which we call the *scaled embedding average*, ensures this linearity. We further analyze other composition functions based on information theory (Amigó et al., 2022). The empirical study in Section 5 verifies the previous theoretical conclusions regarding compositional representations.

Regarding non-compositional representations based on a single token, we observe that in models such as BERT (CLS token) and LLaMA (last token), the growth of information content eventually plateaus, indicating that they do not strictly satisfy linear monotonicity. The CLS token in ModernBERT, however, displays a behavior more consistent with the CER hypothesis.

2 Background

2.1 The Square Norm as Information Content Estimate

Substantial empirical evidence supports the correspondence between the norm of an embedding and the amount of information conveyed by the text (Yokoi et al., 2020; Gao et al., 2019; Schakel and Wilson, 2015; Arefyev et al., 2018; Pagliardini et al., 2018).

At a theoretical level, the analyses by Levy and Goldberg and Arora et al. demonstrate that, in the SGNS framework, the dot product between word vectors approximates the Pointwise Mutual Information (PMI). This result directly implies a correspondence between the amount of information contained in a text and the squared norm of its associated embedding. Being $\vec{w}_x$ the embedding of the sequence x:

$$\langle \vec{w}_x, \vec{w}_x \rangle = \|\vec{w}_x\|^2 \simeq \text{PMI}(x, x) = I(x)$$

Arora et al. (2019) offer an explanation of this phenomenon in the context of generative models. However, they rely on the assumption of isotropy, which is known not to hold in practice (Ethayarajh, 2019; Gao et al., 2019; Cai et al., 2021). In parallel, Oyama et al. (2023) established a formal correspondence in the SGNS framework between the squared norm of a word embedding and the KL divergence between the word's co-occurrence distribution and the unigram distribution. The authors also extended this analysis to contextual embeddings, under certain assumptions.

In general, all these studies focus on the word level. In this work, we analyze the correspondence at the word sequence level by identifying more general sufficient formal conditions and by conducting both theoretical and empirical analyses of composition functions that aim to preserve this correspondence.

2.2 The Constant Entropy Rate

From Shannon's early work in the 1950s to the present, the notion of a *Constant Entropy Rate* (CER) has inspired extensive research in (psycho)linguistics. This principle suggests that, in human communication, the rate of entropy remains relatively stable over time. In other words, although the complexity and diversity of vocabulary may vary, speakers adjust their linguistic choices, such as word length or syntactic structure, so that the amount of information transmitted per unit of time remains constant. A direct formal consequence is that the total self-information of a text, defined as $-log(P(x))$, tends to increase linearly with the number of words, at least beyond a certain length. In other words, each additional word contributes, on average, a similar amount of information. However, in practice, the growth of self-information may exhibit an initial non-linear phase and then stabilize toward a linear trend. Moreover, this linearity is not exact, as the rate of growth may gradually decelerate due to increasing redundancy or predictability in longer texts (Genzel and Charniak, 2002).

Beyond linguistic studies (Aylett and Turk, 2004; Florian Jaeger, 2010), this phenomenon has been validated across diverse datasets and probability estimation techniques (Genzel and Charniak, 2002; Meister et al., 2021).

To the best of our knowledge, the extent to which embedding models reflect this property has not been thoroughly investigated. Verma et al. (2023) reexamine the CER hypothesis using GPT-2. The authors find no clear evidence supporting CER when neural language models are employed, in contrast to the patterns previously suggested by n n-gram models. In our work, we revisit the CER hypothesis across different models as well as composition functions over embeddings.

2.3 Token-based Text Embeddings

Text embedding models that do not rely explicitly on compositional techniques often derive sentence-level representations from a single token, memory cell, or global aggregation mechanism. Early approaches primarily employed recurrent architectures, where models such as Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) or bidirectional LSTMs (Peters et al., 2018) encoded sentence representations by processing sequences token by token and extracting the final hidden state. However, these architectures struggled with long-range dependencies due to their sequential nature and susceptibility to the *vanishing gradient* problem.

Between 2015 and 2021, to address these limitations, several alternative representation mechanisms were proposed, most of which were trained at the sentence level. Examples include encoder-decoder architectures like Skip-thought vectors (Kiros et al., 2015), training word embeddings optimized for direct averaging (Siamese CBOW) (Kenter et al., 2016), hybrid models combining recurrent and convolutional layers (Lai et al., 2015), extensions of Seq2Seq embeddings incorporating attention for classification tasks (CoVe) (McCann et al., 2017), sentence embeddings optimized for pairwise similarity (SBERT) (Reimers and Gurevych, 2019), multi-task training strategies (Subramanian et al., 2018), and contrastive learning approaches (Gao et al., 2021; Giorgi et al., 2021; Yang et al., 2021). While these methods enhanced representation quality, they were often tailored for sentence-level similarity tasks rather than for general-purpose text embedding.

In our experiments, we consider widely adopted non-compositional representations: the CLS token used in BERT (Devlin et al., 2019) and ModernBERT, as well as the final token (EOS) (Neelakantan et al., 2022; Wang et al., 2024) in LLaMA (3.1 8B). These approaches represent state-of-the-art token-based strategies for sentence embedding without explicit compositional mechanisms.

2.4 Compositional Text Embedding

Since the introduction of static word embeddings by Mikolov et al. (2013)., various composition functions have been proposed to represent longer text spans. Simple methods such as the global average—i.e., the mean of all word vectors in a sentence—and vector summation have proven to be robust baselines across numerous tasks (Boleda, 2020; Lenci, 2018; Blacoe and Lapata, 2012; Perone et al., 2018; Baroni and Lenci, 2010; Rimell et al., 2016; Czarnowska et al., 2019; Wieting and Gimpel, 2018; Ethayarajh, 2018; Wieting et al., 2016). Gittens et al. (2017) formally show that, under certain conditions, vector summation approximates paraphrasing of word sets.

Word embedding composition is also a core component of neural language models. For instance, in Transformer architectures, the self-attention mechanism constructs token representations by dynamically aggregating contextual information from all positions in the input. More recently, NV-Embed models (Lee et al., 2025) introduce a *latent attention layer* that computes a weighted aggregation of token embeddings into a single fixed-size vector. However, we exclude this approach from our norm-based analysis as it applies an ℓ_2-normalization step immediately after pooling, producing strictly unit-length embeddings. Since every output vector satisfies $\|\mathbf{e}\|_2 = 1$, its squared norm contains no variation and thus cannot serve as a proxy for IC.

A known limitation of additive methods like averaging or summation is that they disregard word order. To address this, Amigó et al. (2022) propose several information-theoretic composition functions that preserve text structure. These are instantiations of the general form $F_{\lambda,\mu}(\vec{w}_1, \vec{w}_2)$ for various λ and μ values:

$$
\begin{aligned}
G_{\lambda,\mu} = &\ \lambda\left(\|\vec{w}_1\|^2 + \|\vec{w}_2\|^2\right) \ - \ \mu\,(\vec{w}_1 \cdot \vec{w}_2), \\
F_{\lambda,\mu}(\vec{w}_1, \vec{w}_2) = &\ \frac{\vec{w}_1 + \vec{w}_2}{\|\vec{w}_1 + \vec{w}_2\|}\,\sqrt{G_{\lambda,\mu}}
\end{aligned}
$$

This function reduces to simple vector summation ($\texttt{F}_{\texttt{Sum}}$) when $(\lambda, \mu) = (1, -2)$, and to pairwise averaging ($\texttt{F}_{\texttt{Avg}}$) when $\lambda = \frac{1}{4}$ and $\mu = -\frac{1}{2}$ (Amigó et al., 2022). Under the assumption of correspondence between squared norm and IC, the $\texttt{F}_{\texttt{Joint}}$ variant ($(\lambda, \mu) = (1, 1)$) captures the idea that the information content of the composition corresponds to the word joint probability ($I(w_1, w_2) = -\log(P(w_1, w_2))$). The variant F_{ind} ($(\lambda, \mu) = (1, 0)$), assumes statistical independence between the components, i.e., $I(w_1, w_2) = -\log(P(w_1) \cdot P(w_2))$. The third variant, F_{inf} with $(\lambda, \mu) = \left(1, \frac{\min(|\vec{w}_1|, |\vec{w}_2|)}{\max(|\vec{w}_1|, |\vec{w}_2|)}\right)$, is designed to satisfy additional formal constraints such as redundant information (same direction and smaller norm) does not affect the composition[1].

In this paper, we analyze both theoretically and empirically the ability of these composition functions to preserve the relationship between IC and the squared norm of embeddings.

3 Sufficient Theoretical Conditions

In this section, we formally derive two sufficient conditions for the correspondence ($I(w) \propto \|\vec{w}\|^2$) regardless of the embedding model (static or contextual) or the mechanism used to represent full texts (CLS token, EOS, or composition functions). Note that we are not claiming that these conditions always hold, but rather that they are sufficient conditions for the correspondence. The purpose of this formal analysis is to examine under which conditions the correspondence at the text level may hold independently of specific architectures and composition functions.

The first condition (see Section 2.1) is that the IC of the represented content must grow according to its embedding norm without jumps or abrupt peaks. We denote as $\vec{w}$ the embedding of the expression $w = (w_1, \ldots, w_n)$.

Property 1 (Information Monotonicity) *There exists a strict monotonic and differentiable function f such that: $I(w) = f(|\vec{w}|)$.*

This property can be grounded in the notion of distributional semantics. If we assume that the empty expression is represented at the origin of the coordinate system (the zero vector), and that in distributional semantics expressions sharing similar contexts tend to be located close to each other in the space, then the distance from the empty text (the origin) depends solely on the number of contexts in which the expression appears, that is, on its IC.

The second sufficient condition is that each dimension represents independent semantic features and therefore independently affects the probability assigned to the embedding. Therefore, the amount of information contributed by each dimension must be additive. This means that the function estimating the amount of information can be decomposed into a sum of functions for each dimension. That is:

Property 2 (Information Additivity) *For each embedding dimension, there exists a differentiable function $g_i : \mathbb{R} \longrightarrow \mathbb{R}_{\geq 0}$ such that $g_i(0) = 0$ and $I(w) = \sum_{i=1}^{n} g_i(\vec{w}_i)$.*

On the basis of the previous properties, we can formally demonstrate that (see the formal proof that can be found in Appendix A):

Theorem 1 (Embedding Inf. Content) *If properties 1 and 2 hold, then the information content of expressions is proportional to the square of the embedding norm:*

$$
I(w) \propto \|\vec{w}\|^2
$$

Essentially, Property 1 states that I must take the form $f(\|\vec{w}\|)$, where f is a strictly increasing and derivable function. Property 2 states that it must also take the form $I(w) = \sum_{i=1}^{n} g_i(\vec{w}_i)$. Consequently, the information content estimator must satisfy:

$$
I(w) = f\left(\sum_{i=1}^{n} \vec{w}_i^2\right) = \sum_{i=1}^{n} g_i(\vec{w}_i)
$$

Therefore, the function I must take the form:

$$
I(w) = c \cdot \sum \vec{w}_i^2 = c \cdot \|\vec{w}\|^2
$$

4 IC Monotonicity of Compositional Functions

In this section, we formally examine the ability of composition functions to preserve the correspondence $I(w) \propto \|\vec{w}\|^2$ as words are added to a text, assuming the correspondence at the token level and the Constant Entropy Rate mentioned in Section 2.2.

[1] In this work, we apply these composition functions sequentially across tokens to preserve the linear structure of the text. Amigó et al. (2022) suggest that the sequential structure achieves similar results than dependency trees or constituency parses.

Model	Embedding Type	Parameters	Embedding Size	Context Window
GloVe (6B)	Static	-	300	-
Word2Vec	Static	-	300	-
BERT-base-uncased	Contextual	110M	768	512 tokens
ModernBERT-base	Contextual	149M	768	8,192 tokens
LLaMA 3.1 8B	Contextual	8B	4,096	128,000 tokens

Table 1: Characteristics of the models used for embedding generation.

4.1 Averaging and Max-Based Functions

We begin by showing analytically that none of the standard composition functions—namely sum, average, or maximum—satisfy CER. However, the parameterized version of the average, which we refer to as the scaled embedding average, does satisfy it.

We can generalize the summation and averaging composition functions as follows. Let $w_1, \ldots, w_n$ be a sequence of tokens, and let $\vec{w}_{i,j}$ denote the j-th dimension of the embedding of token i, with $j = 1, \ldots, d$ and $i = 1, \ldots, n$. Then, the dimension j of the generalized composition function is defined as:

$$\mathtt{F}_\gamma(\vec{w}_1, \ldots, \vec{w}_n)_j = \frac{\sum_{i=1}^n \vec{w}_{i,j}}{n^\gamma}$$

where $\gamma = 0$ and $\gamma = 1$ correspond to the traditional vector summation and averaging, respectively. Therefore, the squared norm of the composed vector is:

$$\|\mathtt{F}_\gamma(\vec{w}_1, \ldots, \vec{w}_n)\|^2 = \sum_{j=1}^d \left(\frac{\sum_{i=1}^n \vec{w}_{i,j}}{n^\gamma} \right)^2$$

Let $\mathbb{E}_i[\vec{w}_{i,j}]$ denote the expected value of dimension j across tokens. Then, for sufficiently long sequences, the squared norm tends to:

$$\|\mathtt{F}_\gamma(\vec{w}_1, \ldots, \vec{w}_n)\|^2 \simeq \sum_{j=1}^d \left(\frac{n \cdot \mathbb{E}_i[\vec{w}_{i,j}]}{n^\gamma} \right)^2$$

$$= n^{2-2\gamma} \sum_{j=1}^d (\mathbb{E}[\vec{w}_{i,j}])^2$$

This means that the squared norm grows linearly with n when $\gamma = \frac{1}{2}$, decreases with n when $\gamma > \frac{1}{2}$ (e.g., standard averaging), and grows superlinearly when $\gamma < \frac{1}{2}$ (e.g., pure summation)[2].

Therefore, for the standard summation ($\gamma = 0$) or averaging ($\gamma = 1$), the Constant Entropy Rate (CER) does not hold. Instead, an intermediate approach is required—one where the sum of embeddings is normalized by the square root of the number of tokens. We refer to this as the *scaled embedding average*:

$$\mathtt{F}_{\gamma=\frac{1}{2}}(\vec{w}_1, \ldots, \vec{w}_n)_j = \frac{\sum_{i=1}^n \vec{w}_{i,j}}{n^{1/2}}$$

An alternative consists of taking the maximum absolute value in each dimension while preserving the sign of the highest-magnitude component. This serves as a composition function that highlights the most dominant signal per dimension (Zhelezniak et al., 2019).

$$\mathtt{F}_{\mathtt{max}}(\vec{w}_1, \ldots, \vec{w}_n)_j = \text{Argmax}_{\vec{w}_{1,j}, \ldots, \vec{w}_{n,j}}(|\vec{w}_{i,j}|)$$

By definition, this function will exhibit decelerated growth as more tokens are added, and therefore the CER does not hold.

4.2 Information Theory based Composition Functions

Following Amigó et al. (2022), several embedding composition functions based on information theory are proposed. In this section, we show that not all of these functions satisfy CER.

These are specific instantiations of a general composition function denoted as $\mathtt{F}_{\lambda,\mu}(\vec{w}_1, \vec{w}_2)$, defined as:

$$\frac{\vec{w}_1 + \vec{w}_2}{\|\vec{w}_1 + \vec{w}_2\|} \cdot \sqrt{\lambda(\|\vec{w}_1\|^2 + \|\vec{w}_2\|^2) - \mu\langle \vec{w}_1, \vec{w}_2 \rangle}$$

The sum ($\lambda = 1, \mu = -2$) and averaging ($\lambda = 1/4, \mu = -1/2$) parameter instantiations were analyzed in the previous subsection.

Beyond these basic cases, $\mathtt{F}_{\mathtt{Ind}}$ ($\lambda = 1$ and $\mu = 0$) assumes word independence. Its squared norm is additive and leads to a linear growth in information content, satisfying CER.

$$\|\mathtt{F}_{\mathtt{Ind}}(\vec{w}_1, \vec{w}_2)\|^2 = \|\vec{w}_1\|^2 + \|\vec{w}_2\|^2$$

On the other hand, $\mathtt{F}_{\mathtt{Joint}}$ is defined by setting $\lambda = 1$ and $\mu = 1$. The resulting squared norm $\|\mathtt{F}_{\mathtt{Joint}}(\vec{w}_1, \vec{w}_2)\|^2$ is:

$$\|\vec{w}_1\|^2 + \|\vec{w}_2\|^2 - \cos(\vec{w}_1, \vec{w}_2)\|\vec{w}_1\|\|\vec{w}_2\|$$

[2] Linearity in n means $\|\mathtt{F}_\gamma\|^2 \propto n^1$. Since $\|\mathtt{F}_\gamma\|^2 \sim n^{2-2\gamma}$, we require $2 - 2\gamma = 1$, hence $\gamma = \frac{1}{2}$.

In this case, the information content grows linearly only when the embeddings are orthogonal (i.e., $\cos(\vec{w}_1, \vec{w}_2) = 0$), indicating independence. However, if the cosine similarity is greater than $\frac{1}{\|\vec{w}_2\|^2}$, then the information content in the composition decreases (i.e., $\|\mathtt{F_{Joint}}\|^2 < \|\vec{w}_1\|^2$), violating CER.

Finally, $\mathtt{F_{Inf}}$ corresponds to the parameterization $\lambda = 1$ and $\mu = \frac{\min(\|\vec{w}_1\|, \|\vec{w}_2\|)}{\max(\|\vec{w}_1\|, \|\vec{w}_2\|)}$. The intuition is that adding redundant information (same direction, smaller norm) should not alter the original embedding[3]. The squared norm of the composition depends on the redundancy between the embeddings, as captured by their cosine similarity. Being $\|\vec{w}_1\| > \|\vec{w}_2\|$ (see proof in Appendix B):

$$\|\mathtt{F_{Inf}}(\vec{w}_1, \vec{w}_2)\|^2 = \|\vec{w}_1\|^2 + (1 - \cos(\vec{w}_1, \vec{w}_2))\|\vec{w}_2\|^2$$

While $\cos(\vec{w}_1, \vec{w}_2) \leq 1$, this function always yields an increase in norm. However, if the redundancy introduced by composition remains approximately constant, the information content will grow linearly with the length of the expression, thereby satisfying CER.

In summary, our analysis shows that the composition functions that satisfy CER are the variants $\mathtt{F_{Ind}}$ and $\mathtt{F_{Inf}}$ proposed in (Amigó et al., 2022), as well as the scaled $\mathtt{F}_{\gamma=\frac{1}{2}}$ introduced in this paper.

5 Experiments

In this section, we empirically examine the correspondence between the squared norm and the amount of information. At the lexical level, we analyze its correlation with word probabilities in a corpus. At the sequence level, we study how the squared norm grows as tokens are added to a sequence.

5.1 Embedding Models

In our experiments, we use a range of models spanning both *static* and *contextual* representations, as summarized in Table 1. Static embeddings encode words as fixed-dimensional vector representations derived from word co-occurence statistics in large corpora. In this category, we use GloVe (6B) (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013), which produce 300-dimensional embeddings that remain unchanged across different contexts.

For contextual representations, we study transformer-based models that dynamically adjust token embeddings based on surrounding text. Due to their popularity, we consider BERT-base-uncased (Devlin et al., 2019) and the recent model ModernBERT-base (Warner et al., 2024). Both are bidirectional encoders that incorporate both left and right context, generating 768-dimensional token representations with a context window of up to 512 and 8,192 tokens, respectively. ModernBERT employs rotary positional embeddings (RoPE), which can influence the encoding of information content, and its extended context window enables the analysis of significantly longer sequences compared to BERT.

In addition, we consider LLaMA 3.1 8B (Touvron et al., 2023), which is a decoder-based model that computes embeddings autoregressively, conditioning each token's representation on previously processed tokens within a context window of up to 128K tokens. These models provide complementary perspectives on representation learning, allowing us to analyze how embedding norms evolve with text length across different architectures.

5.2 Experiments at the Word Level

To evaluate the correspondence at the lexical level, we computed the information content $(-log(P(w)))$ of the 5,000 most frequent words in the Brown corpus. Additionally, we calculated the squared norm of the corresponding embeddings using different models. Overall, the initial correlations obtained were very low, in contrast to previous empirical studies (Oyama et al., 2023).

Upon inspecting the data, we found that these discrepancies were due to differences in the distribution of words in the Brown corpus compared to the pre-training corpora of the models. In particular, some words that are infrequent in Brown for circumstantial reasons—such as named entities—are not generally rare. For this reason, we applied a smoothing procedure. We divided the 5,000 words into subsets based on intervals of information content $(-log(P(w)))$ in Brown (e.g., from 4 to 5, from 5 to 6, etc.). Then, within each subset, we computed the average information content in Brown and the average squared norm of the embeddings. Finally, we calculated the Pearson correlation over these aggregated points.

Additionally, in order to connect the results with the theoretical analysis, we empirically examined the second sufficient condition, namely, the statis-

[3] When the input vectors are aligned, the function returns the longer one. If $\cos(\vec{w}_1, \vec{w}_2) = 1$ and $\|\vec{w}_1\| > \|\vec{w}_2\|$, then $\mathtt{F_{Inf}}(\vec{w}_1, \vec{w}_2) = \vec{w}_1$.

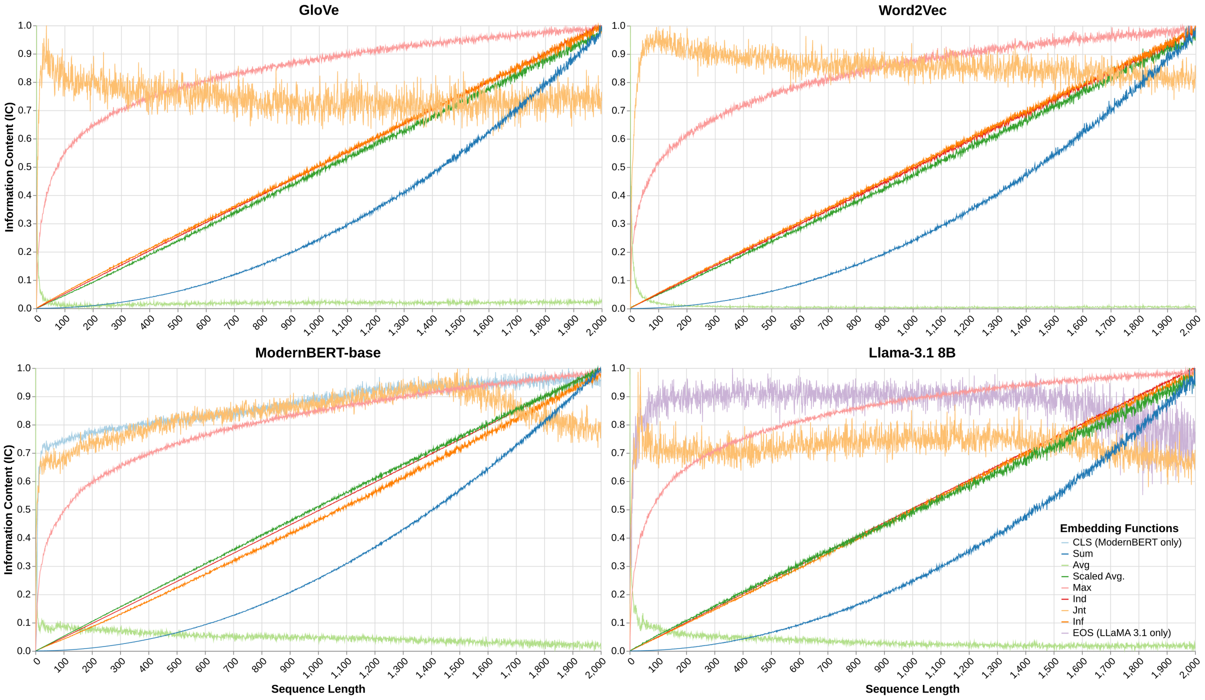

Figure 1: Expected squared norm of embeddings for 1–2000-token sequences: static GloVe and Word2Vec, and transformer last-layer embeddings from ModernBERT-base and LLaMA. The CLS and EOS tokens are reported only for ModernBERT-base and LLaMA, respectively, as these are unique to the architecture of each model.

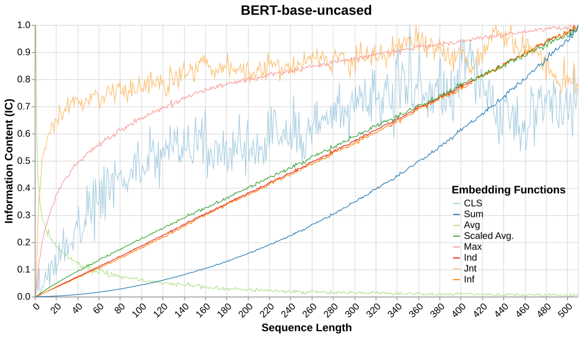

Figure 2: The expected squared norm (amount of information) of last-layer embeddings with a length ranging from 1 to 510 tokens in BERT.

tical independence of embedding dimensions. To this end, we computed the average of the absolute Pearson correlation between dimensions across the 5,000 selected words.

Table 2 presents the results. Since contextual models represent words at the subtoken level, we include in the table only the best-performing composition function for each model. The results show that, in general, models achieve low average absolute dimension correlation—especially static models—indicating statistical independence across dimensions. On the other hand, all models exhibit a strong correlation between IC and squared norm when certain composition functions are applied.

Although not shown in the table, for BERT-base the correlation between IC and norm significantly decreases when using composition functions instead of the CLS token. In contrast, for ModernBERT, the correlation remains stable across composition functions and decreases when using CLS. In the case of LLaMA, correlations remain above 0.95 regardless of the composition function applied.

5.3 Experiment at Sequence Level

In the following experiment, we analyze the monotonicity of the squared norm as words are added to a sequence. Theoretically, under the CER assumption, this growth should be linear, or at the very least monotonically increasing.

For this purpose, we generate embeddings from word sequences in the C4 dataset[4], specifically the *train* split of the *realnewslike* subset from Hugging Face's repository. This corpus consists of filtered English web text, selected to provide a diverse and representative sample of contemporary written language. Text sequences of varying lengths are randomly sampled from the corpus[5]. We consider 100

[4] https://huggingface.co/datasets/allenai/c4

[5] Sequence lengths are determined post-tokenization, ensuring that special tokens are excluded from token counting but included during model inference

Model	Token composition	Avgerage Absolute Dimension Correlation	Correlation IC vs Norm^2
Word2Vec	-	0.045	0.997
GloVe	-	0.053	0.989
bert-base-uncased	CLS	0.125	0.985
ModernBERT-base	$\texttt{F}_{\texttt{max}}$	0.127	0.985
LLaMA 3.1 8B	$\texttt{F}_{\texttt{Joint}}$	0.064	0.971

Table 2: Average absolute dimension correlation and correlation between Information Content (IC) and squared norm across different embedding models at word level.

different sequences for each sequence length (ranging from 1 to 2000 tokens), resulting in a total of 200K samples.

Figure 1 shows the expected squared norm (amount of information) of embeddings with a length ranging from 1 to 2,000 tokens. We observed significant variance in the absolute norms of embeddings depending on the model and composition function. For clarity, we normalized the values for each embedding method between the minimum and maximum across all sequence lengths.

The results for BERT are shown in Figure 2, where the sequences do not exceed 510 tokens. We limited the sequence length in the case of BERT to compare the relative behavior of the embedding based on the CLS token. Since the model was trained with two special tokens in the input, the effective usable sequence length is restricted to 510 tokens.

Averaging based Text Embeddings. The behaviour of sum, average, and scaled average, aligns with our theoretical analysis[6]. For all models, the average tends to decrease as text length increases (light green). The sum composition function exhibits accelerated growth behavior (dark blue line). However, the scaled average behaves linearly in all cases (dark green line).

Maximum Absolute Value Composition. It is denoted as *max* in the figure legend, light pink. It presents a strictly increasing but decelerating behavior. Although it appears to saturate at 2,000 tokens, this is not actually the case. The theoretical square norms given the value ranges are much higher in all models. Therefore, there is still significant room for growth.

Information Theory based Composition. The behavior of $\texttt{F}_{\texttt{Ind}}$, $\texttt{F}_{\texttt{Inf}}$ and $\texttt{F}_{\texttt{Joint}}$ also aligns with the theoretical analysis described in the previous section. $\texttt{F}_{\texttt{Ind}}$ results in an average linear growth as text length increases (red line in the figure). $\texttt{F}_{\texttt{Inf}}$ is also linear, suggesting that information redundancy captured by the cosine component in $\texttt{F}_{\texttt{Inf}}$, is uniformly distributed across texts (orange line). The function $\texttt{F}_{\texttt{Joint}}$ (yellow line) presents a decreasing behavior in some ranges and models[7]. According to Amigó et al. (2020), $\texttt{F}_{\texttt{Joint}}$ should approximate the amount of joint information $-\log(P(x,y))$, so it should be increasing. However, the correspondence proposed by the author assumes an equivalence (not proportionality) between the dot product and the PMI. The composition function $\texttt{F}_{\texttt{Inf}}$ is actually a parameterization of $\texttt{F}_{\texttt{Joint}}$ that solves this issue, being monotonically increasing.

Token based Text Embeddings. The CLS token of BERT exhibits growth that progressively decelerates (light blue line). The results suggest that the growth stops after the first 300 to 400 tokens. We hypothesize that because over 90% of BERT's pre-training sequences are capped at 128 tokens (with the remaining 10% extending to its 512 token-maximum), the model seldom encounters longer contexts and thus has limited opportunity to increase CLS norms beyond that window—explaining the observed breakpoint in linearity. This finding indicates that the CLS token is not suitable for capturing the increase in information quantity in long texts.

In the case of the EOS token in LLaMA, the growth levels off around token 100, showing limitations similar to those of BERT's CLS token. In contrast, the CLS token in ModernBERT displays behavior more in line with the CER hypothesis: self-information increases sharply during the initial tokens and then follows a roughly linear trend, with a gradual deceleration as the sequence lengthens.

Additionally, we have studied the squared norm of individual tokens within the sequence in contextual models. We wanted to analyze the effect of context sequence length on the amount of in-

[6] The small variations are due to the effect of sequence sampling.

[7] We have no concrete explanation for the decrease in information content in ModernBert after token 1500

formation encoded in lexical embeddings. We did not find any consistent pattern across contextual models.

6 Conclusions

Theoretical analyses and empirical findings in the literature support a correspondence between the square norm of a word embedding and the IC of the word it represents. In this work, we extend this correspondence to representations of word sequences. At the theoretical level, we identify two sufficient properties of the representation system that establish this correspondence: the growth of the norm with the IC of the represented text and the statistical independence between dimensions. Although these properties are not strictly necessary, our word-level experiments show that when the appropriate subword composition function is applied, the models satisfy the sufficient conditions and exhibit a strong correspondence between IC and squared norm.

In addition, we have formally analyzed existing composition functions under the assumption that information content increases linearly with text length (Constant Entropy Rate). According to our analysis, validated by experiments, neither summation nor averaging satisfies this property. As a solution, we propose a scaled embedding average, which simply divides by the square root of the number of embeddings being averaged. The analysis also shows that the function $\mathtt{F_{Inf}}$ is able to preserve this property, while max-based composition does not exhibit linear behavior.

Our experiments also suggest that some token-based text embeddings such as CLS in BERT or EOS in LLaMA, do not preserve this property either. However, the CLS token in ModernBERT exhibits a more expected behavior in terms of the CER hypothesis.

This work opens new avenues for studying the representation power of embeddings from a formal perspective and exploring/exploiting the geometric properties of dense text representations in distributional semantics from the perspective of Information Theory.

Limitations

We acknowledge certain limitations in this work, though addressing them in detail would require more space. First, many of the non-compositional approaches discussed in Section 2.3 are trained specifically for sentence representation and are not widely adopted. As a result, we excluded them from our experiments. However, we do not rule out the possibility that these embeddings could exhibit interesting behavior at longer text lengths.

Second, while both the literature and our analysis have shown that the correspondence between IC and the squared norm of embeddings naturally emerges in models based on distributional semantics, some models may still exhibit consistent behavior, provided that specific estimation functions for information content are explicitly defined for them.

Third, the datasets used in our initial word-level correlation experiments between IC and the squared norm are limited. Additionally, more sophisticated methods could be applied to analyze statistical independence across dimensions. As this work focuses on composition functions, we leave a more in-depth investigation of word-level aspects for future research.

Fourth, beyond examining the CER hypothesis, this work could be extended with experiments on the correspondence between the norm and empirical probabilities of word sequences. However, we have not pursued this experimental line due to the difficulty of objectively estimating the probability of long sequences in a corpus.

Fifth, we have not conducted experiments to examine the effect of the correspondence between IC and embedding norm on downstream tasks. We leave this line of research for future work.

Finally, rather than a limitation, a potential avenue for future work is the parametrization of composition functions by empirically analyzing the actual entropy growth rate across a text collection. This could enable the adjustment of composition functions—or even the embedding process itself—to align with this empirical parameter.

Acknowledgments

This work is partially funded by the Spanish Ministry of Science, Innovation and Universities (project FairTransNLP PID2021-124361OB-C32) funded by MCIN/AEI/10.13039/501100011033, by a research agreement between UNED and Red.es (C039/21-OT-AD2) and by ERDF, EU A way of making Europe. Alejandro Benito-Santos' work was supported by grant JDC2022-048408-I funded by MICIU/AEI/10.13039/501100011033 and by

"European Union NextGenerationEU/PRTR". However, the views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Enrique Amigó, Alejandro Ariza-Casabona, Victor Fresno, and M. Antònia Martí. 2022. Information theory–based compositional distributional semantics. *Computational Linguistics*, 48(4):907–948.

Enrique Amigó, Fernando Giner, Julio Gonzalo, and Felisa Verdejo. 2020. On the foundations of similarity in information access. *Information Retrieval Journal*, 23(3):216–254.

Nikolay Arefyev, Pavel Ermolaev, and Alexander Panchenko. 2018. How much does a word weigh? weighting word embeddings for word sense induction. *ArXiv*, abs/1805.09209.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2019. A latent variable model approach to pmi-based word embeddings.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47:31–56.

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

William Blacoe 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.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

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

Paula Czarnowska, 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.

Jacob Devlin, 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 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, Melbourne, Australia. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 EMNLP-IJCNLP*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *ICLR*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. 2017. Skip-gram - {Z}ipf + uniform = vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tom Kenter, 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.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171.

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

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and 6 others. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. 2023. Norm of word embedding encodes information gain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2108–2130. Association for Computational Linguistics.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Matthew E. Peters, 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, New Orleans. ACL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. Measuring word significance using distributed representations of words. *ArXiv*, abs/1508.02297.

Claude Elwood Shannon. 1951. Prediction and entropy of printed english. *Bell System Technical Journal*, 30:50–64.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *Preprint*, arXiv:1804.00079.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vivek Verma, Nicholas Tomlin, and Dan Klein. 2023. Revisiting entropy rate constancy in text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15537–15549, Singapore. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. *Preprint*, arXiv:1511.08198.

John Wieting 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, Australia. Association for Computational Linguistics.

Haoran Yang, Wai Lam, and Piji Li. 2021. Contrastive representation learning for exemplar-guided paraphrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4754–4761, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Francesco Moramarco, Jack Flann, and Nils Y. Hammerla. 2019. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. *Preprint*, arXiv:1904.13264.

7 Appendix

Appendix A: Formal Proof for Theorem 1

We want to prove that if the following equation verifies being f a strictly increasing differentiable function

$$I(\vec{v}) = f\left(\sum_{i=1}^{n} v_i^2\right) = \sum_{i=1}^{n} g_i(v_i)$$

Then, holds

$$I(\vec{v}) = c \cdot \sum_{i=1}^{n} v_i^2 = c \cdot \|\vec{v}\|^2$$

Proof:

I is differentiable because it is the composition of the function f and the square norm, and both are differentiable.

For the proof, we are going to derive I in two ways, and equalizing we will get that if we want to achieve the equality, we will get $\forall x \in \mathbb{R}^+, f'(x) = c$ and therefore the final result.

First, we are going to derive I decomposing it in the following way: $I(\vec{v}) = f(h(\vec{v}))$ where: $h : \mathbb{R}^n \to \mathbb{R}$ and $h(\vec{v}) = v_1^2 + \ldots + v_n^2$.

The gradient of h is:

$$\nabla h(\vec{v}) = (2v_1, \ldots, 2v_n) = 2 \cdot (v_1, \ldots, v_n)$$

Then, applying the chain rule:

$$dI(\vec{v}) = d(f(h(\vec{v})) = f'(h(\vec{v})) \cdot \nabla h(\vec{v}) = \\ f'(v_1^2 + \ldots + v_n^2) \cdot 2 \cdot (v_1, \ldots, v_n)$$

On the other hand, we have $I(\vec{v}) = \sum_{i=1}^{n} g_i(v_i)$. Deriving the sum of functions we get:

$$dI(\vec{v}) = (g'(v_1), \ldots, g'(v_n))$$

So, we need that $\forall v \in \mathbb{R}^n, \forall i \in \{1, \ldots, n\}$:

$$f'(v_1^2 + \ldots + v_n^2) \cdot 2 \cdot v_i = g_i'(v_i)$$

So, if $f'(v_1^2 + \ldots + v_n^2)$ isn't a constant and depends on v_j for some $j \neq i$, the equality can't be true. Like f is strictly increasing, $f'(v_1^2 + \ldots + v_n^2) > 0$ and therefore $f'(v_1^2 + \ldots + v_n^2) = c, c > 0$ for some constant c. So, $\forall x \in \mathbb{R}, f(x) = c \cdot x$ and then:

$$I(\vec{v}) = f\left(\sum_{i=1}^{n} v_i^2\right) = c \cdot \|\vec{v}\|^2$$

Appendix B: The Magnitude of $\mathtt{F_{Inf}}$

The magnitude square of the composition function $\mathtt{F_{Inf}}$ is:

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

With $\lambda = 1$ and $\mu = \frac{min(\|\vec{v}_1\|, \|\vec{v}_2\|)}{max(\|\vec{v}_1\|, \|\vec{v}_2\|)}$. Given that $\frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|}$ is an unitary vector, we can state that:

$$\|\mathtt{F_{Inf}}(\vec{v}_1, \vec{v}_2)\|^2 = \left(\sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}\right)^2 \\ = \lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle$$

Given that $\lambda = 1$ and $\mu = \frac{min(\|\vec{v}_1\|, \|\vec{v}_2\|)}{max(\|\vec{v}_1\|, \|\vec{v}_2\|)}$, then the previous equation is equivalent to;

$$\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \frac{min(\|\vec{v}_1\|, \|\vec{v}_2\|)}{max(\|\vec{v}_1\|, \|\vec{v}_2\|)}\langle\vec{v}_1, \vec{v}_2\rangle$$

In the case of $\|\vec{v}_1\| > \|\vec{v}_2\|$, we obtain:

$$\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \frac{\|\vec{v}_2\|}{\|\vec{v}_1\|}\langle\vec{v}_1, \vec{v}_2\rangle \\ = \|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \frac{\|\vec{v}_2\|}{\|\vec{v}_1\|} cos(\vec{v}_1, \vec{v}_2)\|\vec{v}_1\|\|\vec{v}_2\| \\ = \|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \|\vec{v}_2\| cos(\vec{v}_1, \vec{v}_2)\|\vec{v}_2\| \\ = \|\vec{v}_1\|^2 + (1 - cos(\vec{v}_1, \vec{v}_2))\|\vec{v}_2\|^2$$

Appendix C: Computational Experiments

All experiments were conducted using a NVIDIA RTX 4090 GPU (24GB VRAM). The total estimated computational cost amounts to 200 GPU hours, with an approximate financial cost of 105€, based on standard electricity and cloud-equivalent pricing. The reported computational budget accounts for model inference and embedding generation across all datasets and experimental conditions.

Appendix D: Licensing of Used Artifacts

In this work, we utilize pre-trained models and datasets from Hugging Face's Model Hub. The artifacts are publicly available under their respective licenses, as specified on their respective Hugging Face pages. Specifically, we ensure that all used models and datasets comply with open-source licenses such as Apache 2.0, MIT, or Creative Commons (CC), as applicable. Users can find detailed licensing information on the Hugging Face repository pages of each artifact.

We acknowledge the importance of respecting these licenses and provide proper citations for all resources used in our study.