Understanding Compositionality in Data Embeddings

Anonymous authors
Paper under double-blind review

Abstract

Embeddings are often difficult for humans to interpret, raising potential safety concerns. To address this, we analyze embeddings from different data structures such as words, sentences, and graphs, and interpret them in an understandable manner. This study investigates the algebraic relations, specifically additive, between pairs of vectors that represent entities known to be similar across a particular feature. To this end, we apply two methods: (1) Correlation-based Compositionality Detection, which measures the correlation between known attributes of objects and their embeddings, and (2) Additive Compositionality Detection, which decomposes embeddings into an additive combination of vectors representing specific attributes. Embeddings are evaluated from various models, layers, and training stages to explore their capacity to encode compositional relationships. Sentence embeddings, for example, can be interpreted as the sum of underlying conceptual components. Similarly, word embeddings can be interpreted as capturing a combination of semantic and morphological information. Graph embeddings in recommender systems reflect the sum of a user's demographic attributes. In all three types of data, the relationships between structured entities are encoded as vector operations in embeddings, with a simple operation such as addition playing a central role in expressing compositionality. Code will be publicly available on GitHub upon acceptance.

1 Introduction

The representation of entities, concepts, and relations as vector embeddings is a foundational technique in machine learning (Mikolov et al., 2013a;b). Despite their broad success, embeddings often lack interpretability. One approach to improving interpretability involves examining the *compositionality* of embeddings: if embeddings can be understood as a function of known, interpretable representations, the information they encode becomes more accessible. Recent studies have explored compositionality in various contexts, including sentence embeddings (Hewitt & Manning, 2019), word embeddings (Mikolov et al., 2013a), and graph embeddings (Bose & Hamilton, 2019), as well as broader aspects of neural network structures related to interpretability (Lepori et al., 2023). In particular, sentence embeddings and graph embeddings have shown evidence of a particularly simple, *additive* form of compositionality (Xu et al., 2023; Guo et al., 2023), though this has been observed primarily in limited scenarios.

Word embeddings map words into continuous vector spaces based on their contextual relationships, capturing semantic analogies like *king* - *man* + *woman* $\approx$ *queen* (Mikolov et al., 2013b). Sentence embedding, performed using neural networks like BERT (Devlin et al., 2018), GPT (Brown, 2020) or Llama (Touvron et al., 2023), extends this idea by creating embeddings for entire sentences rather than individual words. It is again reasonable to ask if they display the compositional property. In this case the analogue of *king* - *man* + *woman* $\approx$ *queen* would be *Cat ate mouse* - *mouse* + *bird* $\approx$ *cat ate bird*. Similarly, an interesting question is how graph embeddings capture relationships, like user preferences in bipartite graphs, and encode demographic features such as age or gender. More generally, embeddings represent objects through a relationship function or measurable features, defining a vector space that encodes relational or attribute-based information.

Motivation Understanding whether embedding vectors can be decomposed into distinct semantic components is a fundamental challenge for interpretability. If these embeddings can be broken down into parts corresponding to linguistic inflection, word composition, or collections of attributes, then simple operations such as vector addition may reliably reveal underlying semantic or structural changes. While extensive research has documented additive properties in word embeddings, less is known about whether sentence and graph embeddings exhibit similar behavior. Our work addresses this gap by investigating the extent to which vector addition reflects genuine semantic composition, both in final embeddings and across the internal structure of Transformer models.

Our Contributions Although the decomposition methods we use were introduced in previous work (Xu et al., 2023; Guo et al., 2023), our current study makes several novel contributions:

- Novel Analysis of Linearity: We rigorously assess how much of the semantic and structural information in embeddings can be explained by additive, linear components. Our analysis quantifies the degree of linearity present in embedding spaces, providing insight into how interpretable features emerge from models that are otherwise highly non-linear.
- Extensive Cross-Domain Experiments: We evaluate a diverse array of embeddings—including static word embeddings (e.g., Word2Vec), sentence embeddings from Transformer-based models (e.g., BERT, GPT, Llama), and knowledge graph embeddings used in recommender systems. Our experiments cover multiple model layers and training stages, revealing how the strength of additive signals varies with model complexity and over time.
- Quantitative Findings: Our results demonstrate, for example, that the strength of additive signals in SBERT embeddings increases by up to 15% in mid-level layers before declining in upper layers where task-specific representations dominate. For graph embeddings, the correlation between user behavior and demographic information at later training stages (300 epochs) is 1.5 times higher than at earlier stages (5 epochs), with both stages significantly outperforming the random baseline. Additionally, our analysis shows that word embeddings can be decomposed into distinct morphological and semantic subcomponents, and that graph embeddings encode user attributes via additive operations.

Building on methods from (Xu et al., 2023) and (Guo et al., 2023), we examine the extent of additive compositionality across a range of embedding types, including sentence, word, and graph embeddings, as well as those derived from large language models. For sentences, we focus on pretrained embeddings from Transformer-based architectures, namely GPT (Brown, 2020), Llama 2 (Touvron et al., 2023), and various BERT models (Reimers & Gurevych, 2019; Sellam et al., 2022). To investigate additive decomposability, we introduce a task in which a sentence, for example, *"Can you find me an Adventure movie playing at AMC NewPark in Newark?"*, is split into its constituent concepts (*location*, *theater_name*, and *genre*). We then assess whether these concepts can be additively composed within the embedding space, evaluating performance across different models, various layers within SBERT, and multiple training stages of BERT.

For words, we analyze pretrained word2vec embeddings (Mikolov et al., 2013b), a static model that remains widely used and offers a contrast to the Transformer-based approach. Word2vec embeddings are well known to decompose morphologically. In this paper, the extent to which this decomposition holds across multiple suffixes is investigated, as shown in the example: *weightlessness* - *less* - *ness* + *y* + *ly* = *weightily*. Guo et al. (2023) showed that knowledge graph embeddings trained on the MovieLens dataset (Harper & Konstan, 2015) could be decomposed into demographic attributes, even without this information being used in training. In this paper, we assess the robustness of this finding across an alternative embedding method, and examine the compositionality of embeddings across training stages.

Findings Our experiments show that simple vector operations capture meaningful relationships across various embedding types. For example, we find that sentence embeddings from SBERT, GPT, and Llama-2-7B can be decomposed into additive components that correspond to core conceptual elements, with additive signals strengthening by up to 15% in mid-level layers before diminishing in higher layers. In the case of word embeddings, models such as Word2Vec can be decomposed into coherent morphological and semantic units, aligning well with external benchmarks like WordNet. Similarly, knowledge graph embeddings trained by different scoring functions on the MovieLens dataset show that demographic attributes such as age and gender are encoded through additive operations, even without explicit attribute labels. These results demonstrate that, despite their non-linear origins, large-scale models retain a quantifiable degree of additive compositional structure, providing promising avenues for enhanced interpretability in representation learning.

1.1 Background and Related Work

The principle of compositionality—that the meaning of a whole can be derived from its parts—has long been central to linguistic theory and distributional semantics. The extent to which embeddings are compositional has been an area of research across multiple domains. Identifying the components of an embedding provides insight into the factors affecting its representation, as does understanding the rules by which those parts are combined.

Unlike earlier studies that focus on a single domain or merely report statistical significance, our work provides a unified, quantitative framework for analyzing additive compositionality across words, sentences, and graphs. By clearly separating our empirical findings from established methods, we offer both a rigorous methodological foundation and new insights into the linear signals within modern embedding spaces.

Compositionality in Sentence Embeddings Research on compositionality in sentence embeddings has examined how models represent lexical and syntactic structures. BERT (Devlin et al., 2018) is not given explicit syntactic trees during training, however its representations capture significant syntactic information (Hewitt & Manning, 2019). Ettinger et al. (2016) and Dasgupta et al. (2018) explore semantic roles, scope, and natural language inference by modifying sentence structures, while Adi et al. (2016) evaluate embeddings using tasks like sentence length and word order. Probing tasks and representational similarity analysis (RSA) have further analyzed how models encode linguistic features, with studies like Lepori & McCoy (2020), Chrupała & Alishahi (2019), and Tenney et al. (2019b) investigating syntax, semantics, and sentence structure in models such as BERT and ELMo. RNNs and Transformers have also been shown to encode symbolic structures effectively (Soulos et al., 2020; Yu & Ettinger, 2020). Xu et al. (2023) show that sentence representations from SBERT, using the [CLS] token, exhibit additive compositionality. In this paper, we extend their findings to more recent language models.

Compositionality in Word Embeddings Numerous approaches have been proposed to understand compositionality in word embeddings. Mikolov et al. (2013a) demonstrated that words can be decomposed semantically and morphologically, for example, *quickly* - *quick* + *slow* = *slowly*. Disentangled Representation Learning (Bengio et al., 2013) separates attributes in embeddings, enhancing interpretability by associating latent dimensions with discrete features. Additive compositionality in skip-gram word vectors has been theoretically justified (Gittens et al., 2017), with subsequent work relaxing these assumptions and proposing models that satisfy more realistic conditions (Seonwoo et al., 2019). Shwartz & Dagan (2019) examined compositionality via six tasks, exploring semantic drift and implicit meaning, while Andreas (2019) proposed a metric based on fidelity in reconstructing representations from primitives. In this paper, we use techniques from Xu et al. (2023) originally designed for the analysis of sentence embeddings, and show that the same techniques can be applied to word embeddings. We show that word embeddings can be decomposed into multiple prefixes and suffixes, extending the analysis from Mikolov et al. (2013b). We further compare word2vec embeddings with WordNet embeddings (Miller, 1995) to evaluate their semantic compositionality.

Compositionality in Graph Embeddings While compositionality has been extensively studied in linguistics, less attention has been given to graph embeddings. Graph embeddings, such as those in knowledge graphs and recommender systems, encode relationships and attributes as constituent components within their vector representations. For instance, embeddings trained from movie ratings capture compositional information, including user preferences and movie characteristics. Recent work has explored how graph embeddings encode such components during training, with methods like adversarial loss used to isolate specific attributes (Bose & Hamilton, 2019; Fisher et al., 2020). In this work, we evaluate compositionality in graph embeddings across different scoring functions and training stages, focusing on how embeddings encode and combine diverse attributes and relationships.

Compositionality in Deep Learning More Generally The ability of neural networks to reason compositionally has been studied across various architectures. Kim & Linzen (2019) compared Gated Recurrent Units (GRUs) and Simple Recurrent Networks (SRNs) for compositional generalization, while Wu et al. (2020) analyzed contextual word representation models across architectures. Research increasingly focuses on improving generalization to unseen compositions, with Lake & Baroni (2018) and Kim & Linzen (2020) highlighting limitations of RNNs, LSTMs, and Transformers in systematic tasks. To address this, Lake & Baroni (2023) proposed a meta-learning approach for systematic generalization. Hupkes et al. (2020) introduced task-independent tests to evaluate compositional generalization, showing differences across recurrent, convolutional, and transformer models. Additionally, structural and functional studies (Mu & Andreas, 2020; Lepori et al., 2023) explore how architectures contribute to or hinder compositional understanding.

2 Data Embedding and Compositionality Detection Methods

2.1 Embedding Techniques

Embedding Words Word embeddings involve building vectors for words based on the distributional hypothesis: words that occur in similar contexts have similar meanings. Word embeddings can be divided into contextual (Peters et al., 2018; Devlin et al., 2018) and static (Mikolov et al., 2013b; Pennington et al., 2014) embeddings. Static embeddings model word meanings as one static vector, whereas the embeddings produced by contextual embedding methods differ based on the context in which the word occurs. In this paper we focus on the semantics and morphology of embeddings. Since contextual embeddings naturally additively encode morphology through character level or subword embeddings, we focus instead on static embeddings, specifically pre-trained skip-gram word2vec (Mikolov et al., 2013a).

Embedding Sentences We focus on Transformer-based language models and examine the additive compositionality present in pretrained embeddings. Xu et al. (2023) show that sentence embeddings modelled by the [CLS] token of SBERT can be additively decomposed into the words in the sentence. We extend their results to examine compositionality through different layers of SBERT and through training stages of the MultiBERTs (Sellam et al., 2022). We further examine the additive compositionality present in sentence embeddings from GPT Embeddings and Llama 2. Since these are produced by averaging the word embeddings, we alter the task to instead show that sentence embeddings can be decomposed into a sum of the concepts contained within them. Details are given in Section 3.

Embedding Knowledge Graphs Knowledge Graph Embeddings (KGEs) represent entities and relationships as vectors. Scoring functions over these vectors encode the graph's topology (Nickel et al., 2011). Two main scoring functions are widely used: multiplicative scoring (Yang et al., 2014), which models interactions through the product of entity and relation embeddings, and additive scoring (Bordes et al., 2013), where relationships are modelled as translations in the embedding space. These embeddings are evaluated on link prediction tasks. Guo et al. (2023) previously showed that KGEs trained on the MovieLens dataset using the multiplicative scoring function could be additively decomposed into attribute embeddings encoding information such as age or gender. In this paper, we examine whether KGEs trained using the additive scoring function exhibit similar compositionality. We further examine how this compositionality changes over training stages.

2.2 Compositionality Detection Methods

Previous work has explored methods to detect compositionality in embeddings by examining the relationships between entity attributes and their representations. Two approaches are *Correlation-Based Compositionality Detection* and *Additive Compositionality Detection* (Xu et al., 2023; Guo et al., 2023), which we describe briefly here. For each entity—whether user, word, or sentence—we consider two representations:

- A *binary attribute matrix* $\mathbf{A}$ (e.g., demographics or syntactic properties), where each row corresponds to an entity and each column represents an attribute. Entries $a_{ij} \in \{0, 1\}$ indicate the absence or presence of attribute j in entity i.

- A *continuous embedding matrix* $\mathbf{U}$ (e.g., user behaviour embeddings or word embeddings), where each row corresponds to an entity's embedding and each column represents a dimension of the embedding space.

Correlation-Based Compositionality Detection This method uses *Canonical Correlation Analysis (CCA)* to uncover latent correlations between the attributes and embedding dimensions (Shawe-Taylor et al., 2004). Given the binary attribute matrix $\mathbf{A} \in \{0,1\}^{q\times n}$ and the embedding matrix $\mathbf{U} \in \mathbb{R}^{q\times m}$, where q is the number of entities, n is the number of attributes, and m is the dimensionality of the embeddings, the goal is to find transformation matrices $\mathbf{W}_A \in \mathbb{R}^{n\times k}$ and $\mathbf{W}_U \in \mathbb{R}^{m\times k}$ that maximize the correlation between the projected data:

$$\rho = \max_{\mathbf{W}_A,\, \mathbf{W}_U} \operatorname{corr}\left(\mathbf{A}\mathbf{W}_A,\, \mathbf{U}\mathbf{W}_U\right)$$

Here, corr denotes the Pearson correlation coefficient (PCC) between the projected representations, and k is the number of canonical components. The transformations $\mathbf{A}\mathbf{W}_A$ and $\mathbf{U}\mathbf{W}_U$ capture the most correlated aspects of the attributes and embeddings, respectively. We provide a detailed formal exposition of all settings necessary to define a CAA component, including the precise definition, the underlying vector space and mapping of semantic attributes, as well as illustrative examples. A complete derivation and examples are included in the Appendix A.

Additive Compositionality Detection The additive compositionality detection method quantifies how much each attribute influences the embeddings (Xu et al., 2023). It assumes that an entity's embedding can be approximated as a linear combination of attribute embeddings. Given $\mathbf{A}$ and $\mathbf{U}$, the goal is to solve the linear system $\mathbf{A}\mathbf{X} = \mathbf{U}$ where $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the matrix of attribute embeddings to be learned.

A *leave-one-out (LOO) experiment* is performed for each entity i. We firstly exclude the i-th row from $\mathbf{A}$ and $\mathbf{U}$ to obtain $\mathbf{A}_{-i}$ and $\mathbf{U}_{-i}$, then solve $\mathbf{A}_{-i}\mathbf{X} = \mathbf{U}_{-i}$ using the pseudo-inverse to obtain $\mathbf{X}$. We then estimate the left-out embedding using $\hat{\mathbf{u}}_i = \mathbf{a}_i\mathbf{X}$, where $\mathbf{a}_i$ is the i-th row of $\mathbf{A}$, and finally compare $\hat{\mathbf{u}}_i$ with the actual embedding $\mathbf{u}_i$ using (1) L2 loss, (2) cosine similarity, and (3) retrieval accuracy. More details are given in Appendix B.

Hypothesis Testing To determine statistical significance, a non-parametric hypothesis test is performed by directly estimating the p-value through Monte Carlo sampling. First, the test statistic T_{real} is computed for the real pairing: for CCA, this is the canonical correlation ρ_{real}; for additive compositionality, they are L2 loss, cosine similarity and retrieval accuracy based on the real pairing. Next, permuted pairings are generated by randomly shuffling the rows of $\mathbf{A}$ to disrupt the alignment, resulting in permuted datasets $\{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}\}$.

For each permuted pairing j, the test statistic $T^{(j)}_{\text{perm}}$ is calculated, and the p-value is estimated as $p = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(T^{(j)}_{\text{perm}} \geq T_{\text{real}})$, where $\mathbb{I}(\cdot)$ is the indicator function. A low p-value indicates that the observed statistic is unlikely under random pairings, supporting the statistical significance of the real pairing.

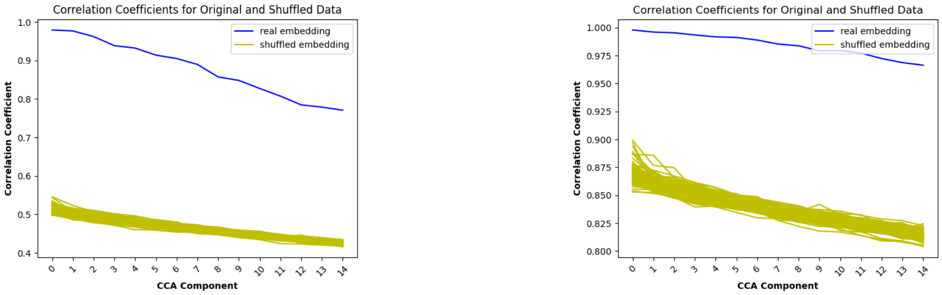
3 Evaluating Compositionality of Data Embeddings

3.1 Sentence Embeddings

We evaluate the extent to which sentence embeddings from SBERT, GPT, and Llama can be additively decomposed into the concepts expressed in the sentence, using a dataset derived from the Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020)[1]. We firstly look at the compositionality of sentence embeddings from the final layer of each model. We go on to examine the compositionality of sentence embeddings through different layers of SBERT, and finally look at how compositionality develops during the training stages of the MultiBERTs.

---

[1] https://github.com/google-research-datasets/dstc8-schema-guided-dialogue



(a) SBERT embeddings

(b) GPT Embeddings

Figure 1: Pearson's correlation coefficient (PCC) for true sentence-concept pairings and 100 permuted pairings. For both SBERT (Figure 1a) and GPT (Figure 1b) embeddings, we see that the correlation of true pairs is significantly higher than for random pairings.

3.1.1 Dataset

We build a dataset consisting of sentences annotated with concepts, taken from the Schema-Guided Dialogue (SGD) dataset. An example pair of sentence and concept labels is:

Sentence: Can you find me an Adventure movie playing at AMC NewPark in Newark?

Concepts: [*location*, *theater name*, *genre*].

We select 2,458 sentences, each annotated with minimum 3 and maximum 4 concepts from a total set of 47 concepts. The mean number of concepts per sentence is 3.16. There are 90 unique combinations of concepts used. The sentences used comprise the test set of the Schema-Guided Dialogue (SGD) dataset. This split was chosen because it is annotated with denser labels, providing more comprehensive concept coverage.

3.1.2 Experiments and Results

Final Layer Embeddings We generate embeddings using pretrained sentence models, producing matrices $\mathbf{U} \in \mathbb{R}^{2,458\times d}$, where d is the dimensionality of the model. For SBERT[2] and Llama[3], embeddings are generated by mean pooling over the token embeddings, excluding padding tokens. For GPT[4], sentence embeddings are obtained from the OpenAI API, which provides precomputed embeddings representing the entire sentence. Each row of the matrix $\mathbf{U}$ consists of a sentence embedding for the corresponding sentence. We construct attribute matrices $\mathbf{A} \in \{0,1\}^{2,458\times 47}$ with each row being a binary vector indexing the relevant concepts for the corresponding sentence.

We assess the compositionality of the sentence embeddings using the methods described in Section 2.2, and the $\mathbf{U}$ and $\mathbf{A}$ matrices described above. For the correlation-based compositionality experiment, all sentences are used to compute correlations. For the additive compositionality experiments, sentences are grouped by their concept combinations, and mean embeddings are computed for each group.

For SBERT and GPT embeddings[5], we apply correlation-based compositionality detection to assess whether sentence embeddings correlate with their binary concept vectors. Figure 1 shows significant difference in correlation scores between real and permuted pairings (p-value < 0.01) for both GPT and SBERT embeddings, indicating that sentence embeddings are correlated with their binary concept vectors.

Results from the additive compositionality detection experiment are reported in Table 1. Cosine similarity and retrieval accuracy for the true pairings are all significantly and substantially higher than for the permuted pairings. The behaviour of SBERT and GPT embeddings are very similar, whereas Llama embeddings

---

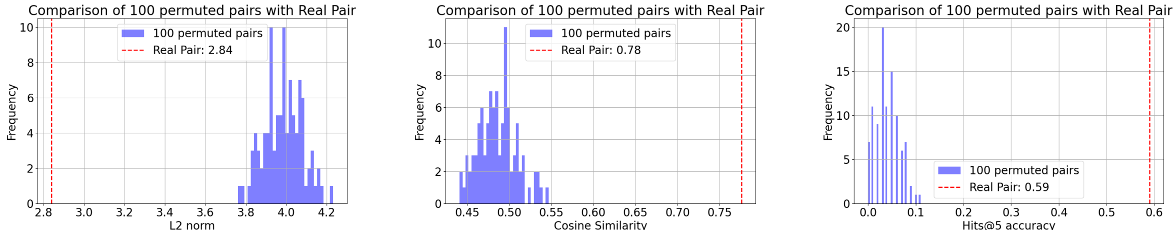
[2] sentence-transformers/all-MiniLM-L6-v2

[3] meta-llama/Llama-2-7b-hf

[4] text-embedding-3-small

[5] We do not assess correlation for Llama embeddings as the dimensionality of these embeddings is too high

have a less substantial increase in compositionality over permuted pairings, as well as a lower retrieval accuracy (Hits@5). Plots of the distribution of the metric values for permuted pairs vs. real pairs for SBERT embeddings are given in Figure 2. Results for GPT and Llama embedding are provided in Appendix D, Figures 9 and 10.



(a) Linear System Loss (b) Cosine Similarity (c) Retrieval Accuracy@5

Figure 2: Test statistics for SBERT's embedding decomposition. Dashed line is the average performance of $\hat{\mathbf{u}}$ learned from the sentence embedding. Bars are the distribution of the results from 100 random permutations. (a) L2 loss, (b) Cosine Similarity, and (c) Retrieval Accuracy@5 compare real SBERT embedding pairs to permuted pairs.

Table 1: Additive compositionality metrics for SBERT, GPT, and Llama embeddings. Figures in **Real** columns are the mean across all leave-one out experiments. Figures in **Permuted** columns are the mean across 100 permutations of the sentence-concept pairs. All increases of Real over Permuted values are significant with $p<0.01$.

Embedding	**Cosine Real**	**Cosine Permuted**	**Hits@5 Real**	**Hits@5 Permuted**
SBERT	0.7761	0.4865	0.59	0.0405
GPT	0.7753	0.4941	0.57	0.0399
Llama	0.9355	0.8153	0.52	0.0468

Comparison between Different Layers We assess the additive compositionality of sentence embeddings derived from each layer of SBERT. At each layer of SBERT, the binary attribute matrix $\mathbf{A}$ remains the same as in the final layer experiment. Separate continuous embedding matrices $\mathbf{U}_i$ are built for each layer i. To generate sentence embeddings at each layer, we extract hidden states for all layers from the SBERT model and apply the same pooling method as used for the last layer (mean pooling over token embeddings). Each embedding is normalized to length 1.

The results, reported in Table 2, show that compositionality increases through the layers, peaking at layer 4 or 5. However, an abrupt drop in compositionality, as measured by cosine similarity, is observed at the last layer. This is in line with the phenomenon that semantic information is better encoded at earlier layers in the model, see for example experiments in the original BERT paper (Devlin et al., 2018). The fact that the cosine metric is fairly high for the random baseline. However, the discrimination between sentence meanings is still high as can be seen by fact that the Hits@5 for real pairings is consistently substantially higher than that for for permuted pairings.

Comparison between Different Training Stages We further compare additive compositionality across different training stages of the MultiBERTs (Sellam et al., 2022). We again use the same binary attribute matrix $\mathbf{A}$ at each training stage. We build continuous embedding matrices $\mathbf{U}_{steps}$ for 0, 20k, 40k, 100k, 1000k, and 2000k training steps. We use the [CLS] token to represent sentences. At 0k steps, there is no significant differences between the values of the cosine similarity metrics for real and permuted despite high overall values (0.9884). This indicates that the embeddings lack differentiation and do not capture conceptual relationships. After 20k training steps, compositionality metrics significantly improve (Cosine Rel. Diff in Table 3), implying that additive compositionality emerges from training. Early stages (0k to 20k steps) show

Table 2: Additive compositionality metrics across SBERT layers. Figures in **Real** columns are the mean across all leave-one out experiments. Figures in **Permuted** columns are the mean across 100 permutations of the sentence-concept pairs. All increases of Real over Permuted values are significant with $p<0.01$.

Layer	**Cosine Real**	**Cosine Permuted**	**Hits@5 Real**	**Hits@5 Permuted**
0	0.8889	0.7808	0.59	0.0397
1	0.9366	0.8671	0.57	0.0406
2	0.9397	0.8576	0.61	0.0390
3	0.9403	0.8330	0.64	0.0412
4	0.9408	0.8298	0.66	0.0400
5	0.9409	0.8273	0.62	0.0417
6	0.7761	0.4865	0.59	0.0405

gains in cosine similarity and retrieval accuracy, demonstrating the model's ability to represent sentences as combinations of concepts. However, further training produces diminishing returns, as the model may shift focus to other linguistic features or risk overfitting. See Table 3 for details.

Table 3: Additive Compositionality Metrics at Different Training Steps of BERT. Number of training steps is given by the suffix to the model, e.g. cls_20k indicates the performance of the model after 20k training steps. Rel. Diff expresses this difference as a percentage of the permuted similarity, showing proportional improvement. Figures in Real columns are the mean across all leave-one out experiments. Figures in Permuted columns are the mean across 100 permutations of the sentence-concept pairs.

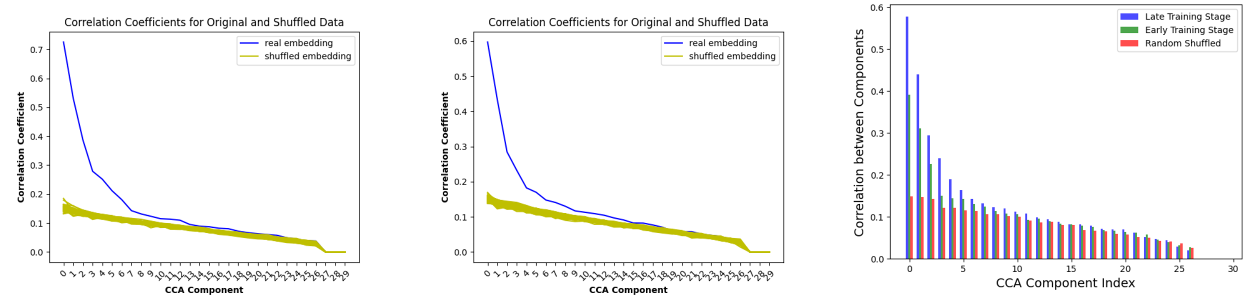
Model	**Cosine Real**	**Cosine Permuted**	**Cosine Rel. Diff**	**Hits@5 Real**	**Hits@5 Permuted**
cls__0k	0.9884	0.9882	0.02%	0.44	0.0418
cls__20k	0.8787	0.7722	13.79%	0.55	0.0407
cls__40k	0.8773	0.7724	13.59%	0.48	0.0405
cls__100k	0.9201	0.8323	10.54%	0.55	0.0417
cls__1000k	0.9545	0.9149	4.33%	0.44	0.0408
cls__2000k	0.9538	0.9094	4.89%	0.48	0.0415

3.1.3 Additive Compositionality in SVO sentences

Xu et al. (2023) demonstrated that the [CLS] embedding in SBERT can be decomposed into its component words, specifically subject, verb, and object (SVO). Building on this foundation, we test the compositionality of syntactic structures across various dimensions, including different training stages, token representations (examining whether individual word embeddings can also be decomposed into SVO components), and layers within SBERT. Results show that the retrieval accuracy of the [CLS] token embedding is significantly, though not substantially, higher than that of any other token embedding in the sentence. Across layers in SBERT, the [CLS] token exhibits the highest additive compositionality across metrics up to layer 10. Across different training stages, a slight decrease in retrieval accuracy is observed, indicating that training minimally reduces additive compositionality. Further details are provided in Appendix F, Table 7, and Figures 12, 13, 14.

3.2 Knowledge Graph Embedding

In Guo et al. (2023), it was shown that knowledge graph embeddings trained using the multiplicative scoring function (Yang et al., 2014) exhibit correlation-based and additive compositionality. In this section, we firstly show that these results also hold for embeddings trained using the additive scoring function (equation 4, Bordes et al. (2013)). We go on to examine the evolution of correlation-based compositionality over training stages, and finally examine the association between canonical variables and user attributes.



(a) Pearson's correlation coefficient (PCC) for true user-attribute pairings and 100 permuted pairings. Embeddings are computed by the additive scoring functions.

(b) Embeddings computed by multiplicative scoring function. Figure recreated from Guo et al. (2023)

(c) Correlation coefficients over different training stages of multiplicative scoring-function-based Knowledge Graph embedding

Figure 3: PCC is calculated between projected $\mathbf{A}$ and projected $\mathbf{U}$. x axis stands for the kth components, y axis gives the value.The PCC value for real pairings is larger than for any permuted pairings.

3.2.1 Compositionality across Diverse KGEs

We use the MovieLens 1M dataset (Harper & Konstan, 2015) with 6040 users, 3900 movies, and 1 million ratings (1–5). As in the sentence embedding experiment, we pair each entity, i.e. user, with two descriptions: a binary vector representing demographic attributes (gender, age and occupation) and a user embedding learned from movie preferences.

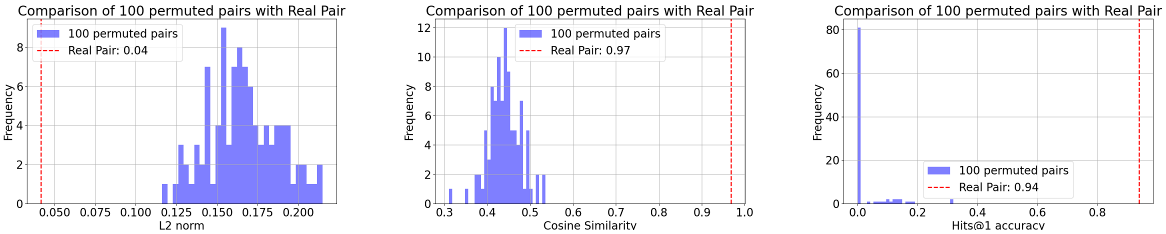
Continuous embeddings are trained as follows. We encode $(user, rating, movie)$ as relational triples (h, r, t) and train 50-dimensional embeddings with OpenKE (Han et al., 2018), using the additive scoring function (equation 4), for 300 training steps. We assess the quality of the embeddings on a link prediction task, obtaining RMSE of 0.92 and Hits@1 of 0.46. Details of the process and evaluation are given in Appendix I.

Correlation-based Compositionality for KGEs The continuous embedding matrix $\mathbf{U}_{corr} \in \mathbb{R}^{6040\times 50}$ for correlation-based compositionality detection experiment is populated with these embeddings. We also generate a binary attribute matrix $\mathbf{A}_{corr} \in \{0,1\}^{6040\times 9}$ containing one 9-dimensional binary attribute vector for each user, including attributes based on gender, age, and occupation.

Additive Compositionality for KGEs For the additive compositionality experiment, we generate a binary attribute matrix $\mathbf{A}_{add} \in \{0,1\}^{14\times 9}$ containing one 9-dimensional binary attribute vector for each user including attributes based on gender and age (2 genders × 7 age groups). The choice of attributes is made to match the choices in Guo et al. (2023). The continuous embedding matrix $\mathbf{U}_{add} \in \mathbb{R}^{14\times 50}$ is generated by taking the mean across embeddings associated with the same attribute combination.

Results Results of the correlation-based compositionality detection method are reported in Figure 3a. Results from Guo et al. (2023) are recreated in Figure 3b for comparison. Compositionality, as measured by the correlation-based compositionality detection method, is shown to be robust across these two different graph embedding methods. The additive scoring-function-based model (Figure 3a) shows a significant difference between real and random pairings, confirming that user embeddings encode demographic information.

Results of the additive compositionality detection experiment are reported in Figure 4, showing L2 loss of 0.04, cosine similarity of 0.97, and Hits@1 of 0.94, each outperforming the random baseline with $p = 0.01$. These findings reject the null hypothesis, demonstrating that user embeddings encode additive relationships between gender and age.



(a) Linear System Loss (b) Cosine Similarity (c) Retrieval Accuracy@1

Figure 4: The test statistics for user additive scoring function embedding decomposition. The dashed line represents the average performance of $\hat{\mathbf{U}}$ from user embeddings, while the bars show the distribution of results from 100 random permutations. (a) L2 loss, (b) Cosine Similarity, and (c) Retrieval Accuracy@5 compare real user embedding pairs to permuted pairs.

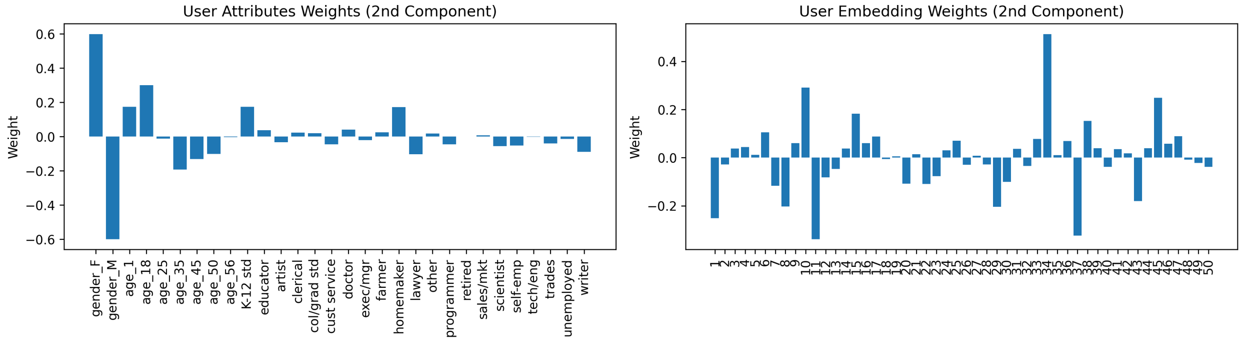


Figure 5: Positive weights in the second canonical component of CCA show the association between user attributes (e.g., *Female* and *Age: 18*) and user embedding dimensions (e.g., *34* and *10*).

3.2.2 Compositionality across Training Stages

We train embeddings for the MovieLens-1M dataset using the multiplicative scoring function (Yang et al. (2014)). Embeddings are extracted at both early (5 epochs) and late (300 epochs) stages of training. We assess the correlation-based compositionality at each training stage. For each stage the attribute matrix $\mathbf{A}$ is the same as $\mathbf{A}_{corr}$ described above, whereas $\mathbf{U}_{early}$ and $\mathbf{U}_{late}$ contain embeddings from the early and late training stages respectively.

Results Results are reported in Figure 3c. We see that as training progresses, the compositionality increases. This means that more demographic information becomes encoded into the embeddings.

3.2.3 Interpretation of Weights

Canonical Correlation Analysis (CCA) identifies linear combinations of variables, called canonical variables, from two datasets (user attributes and user embeddings) that are maximally correlated. The weights in CCA indicate the contribution of each original variable to the canonical variable, with the sign reflecting the direction of the relationship.

In Figure 5, positive weights on the user attribute side (e.g., *Female* and *Age: 18*) and user embedding side (e.g., dimensions *34* and *10*) highlight their strong contribution to their respective canonical variables. This strong correlation between user-side and movie-side canonical variables reveals meaningful links between demographic traits and movie preferences.

3.3 Word Embedding

We apply the same methods to examine two distinct signals contained in word2vec embeddings: semantic and syntactic information. To detect these signals, we use WordNet embeddings as semantic representation, and MorphoLex for syntactic structures. By comparing the word2vec embeddings against both WordNet and MorphoLex, we are able to disentangle the semantic and syntactic aspects of the word2vec representation.

3.3.1 Datasets

WordNet Embeddings WordNet (Miller, 1995) is a large lexical database of English that combines dictionary and thesaurus features with a graph structure. It organizes nouns, verbs, adjectives, and adverbs into synsets (sets of cognitive synonyms) interlinked by semantic and lexical relations. We use WN18RR (Dettmers et al., 2018), a subset of WordNet with 40,943 entities and 11 types of relation.

MorphoLex MorphoLex (Sánchez-Gutiérrez et al., 2018) provides a standardized morphological database derived from the English Lexicon Project, encompassing 68,624 words with nine variables for roots and affixes. An example extract from the dataset is given in Table 4. In this paper, we focus specifically on words with one root and multiple suffixes.

Table 4: Suffix presence (indicated by '1') for selected words from the MorphoLex dataset

Word	er	t	y	est	ly	ness	less
weightier	1	1	1	0	0	0	0
weightiest	0	1	1	1	0	0	0
weightily	0	1	1	0	1	0	0
weightiness	0	1	1	0	0	1	0
weightlessly	0	1	0	0	1	0	1

3.3.2 Experiments and Results

Correlation-based Compositionality of Semantics and Morphology in word2vec We examine the extent to which word2vec embeddings are correlated with WordNet embeddings and with MorphoLex representations. For WordNet, 25,781 words are selected from the intersection of the WN18RR and word2vec (GoogleNews-vectors-negative300) vocabularies. We train embeddings for these words over WordNet on an entity prediction task, which involves predicting the tail entity given a head entity and relation. Training details are provided in Appendix G.1. We build a continuous embedding matrix $\mathbf{U}_{w2v-wn} \in \mathbb{R}^{25,781\times 300}$ using vectors from GoogleNews-vectors-negative300. We build a second continuous embedding matrix $\mathbf{U}_{wn-w2v} \in \mathbb{R}^{25,781\times 20}$ and assess the correlation using the methods described in section 2.

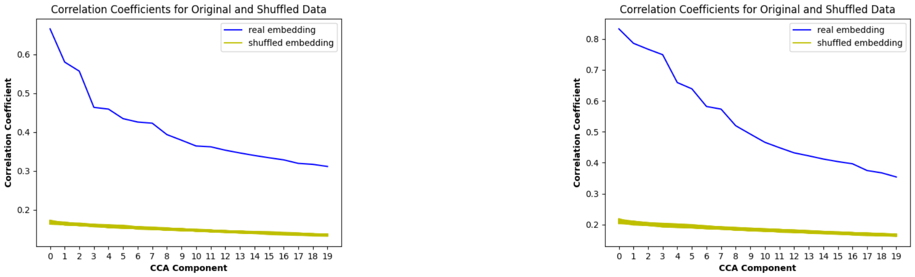
To assess the correlation between word2vec and MorphoLex, 15,342 words and 81 suffixes are selected by intersecting the MorphoLex and word2vec (GoogleNews-vectors-negative300) vocabularies and filtering out suffixes occurring fewer than 10 times. We build a continuous embedding matrix $\mathbf{U}_{w2v-morpho} \in \mathbb{R}^{15,342\times 300}$ using vectors from GoogleNews-vectors-negative300 and a binary embedding matrix $\mathbf{X}_{morpho} \in \{0,1\}^{15,342\times 81}$. We again assess the correlation using the methods described in section 2.

As shown in Figures 6a and 6b, the Pearson Correlation Coefficient (PCC) between word2vec, WordNet embeddings, and MorphoLex binary vectors significantly exceeds the randomized baseline, indicating that word2vec embeddings, trained on contextual co-occurrences, implicitly capture both semantic and morphological information.

Additive Compositionality of word2vec Embeddings To examine the additive compositionality of word2vec embeddings across multiple suffixes, we select 278 words as follows. We filter words that have exactly 3 suffixes, each of which occurs 10 times or more[6]. This results in 17 suffixes, but some words end up

---

[6] These words are listed in Appendix H

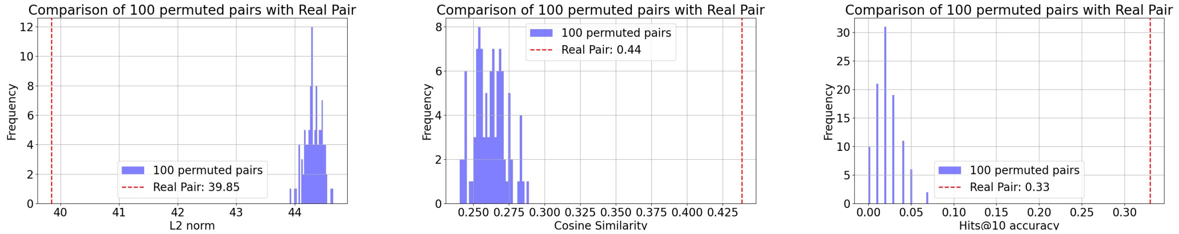


(a) PCC for the true WordNet-word2vec pairings and 50 permuted pairings.

(b) PCC comparison for the true MorphoLex-word2vec pairings and 50 permuted pairings.

Figure 6: PCC is calculated between projected $\mathbf{A}$ and projected $\mathbf{U}$. The first 20 components are selected for illustration. Real pairings show higher PCC values than permuted ones, showing word2vec embeddings capture both semantic and morphological information.

with fewer than 3 suffixes because certain suffixes occur less than 10 times and are filtered out. We build a continuous embedding matrix $\mathbf{U}_{w2v-add} \in \mathbb{R}^{278\times 300}$ using vectors from GoogleNews-vectors-negative300. We build a binary attribute matrix $\mathbf{A}_{morpho-add} \in \mathbb{R}^{278\times 45}$ indicating the presence or absence of each root word and morphemes. We perform the additive compositionality experiment described in section 2 to determine whether embeddings may be decomposed into multiple suffixes. Results are presented in Figure 7. We see that word2vec embeddings can be decomposed into root and multiple suffixes fairly well. The linear system loss is 39.85, lower than the minimum loss of the random system (43.91). Cosine similarity is 0.44, greater than all instances of the random baseline, and retrieval accuracy @ 10 is greater than that of the random system. However, overall these values are low, showing that there is still a fair bit of information that is not being captured by this representation.



(a) Linear System Loss

(b) Cosine Similarity

(c) Retrieval Accuracy@10

Figure 7: The test statistics for word2vec embedding decomposition. Dash line is the average performance of $\hat{\mathbf{u}}$ learned from the word2vec embedding. The bars are the distribution of the results from random permutations that run for 100 times. (a) L2 loss, (b) Cosine Similarity, and (c) Retrieval Accuracy@10 compare real word2vec embedding pairs to permuted pairs.

4 Discussion and Conclusion

In many cases, embeddings are not easily explainable to humans, which may present safety concerns. To address this issue, we analyzed embeddings from word, sentence, and graph data structures and worked to interpret them in a more understandable way. Our central goal was to determine whether embeddings representing structured entities could be decomposed through additive composition. To this end, we applied two methods from the literature: (1) a correlation-based compositionality analysis that measured how well learned embeddings correlated with known attributes, and (2) an additive compositionality detection method (Xu et al., 2023), which provides a way to decompose embeddings into vectors representing distinct attributes.

Our experiments demonstrated that embeddings across diverse domains exhibit additive compositionality, albeit to varying degrees. In word embeddings (e.g., word2vec), morphological information showed a weaker additive signal than expected. For sentence embeddings, we evaluated two types of additive compositionality: (1) decomposition into subject, verb, and object (SVO) and (2) decomposition into constituent concepts. The simpler SVO experiment shows that even a weighted sum of residuals, combined through an MLP, retains notable additive properties. However, decomposing sentences into multiple concepts is inherently more complex due to interactions within the token representations, making this a more stringent test of additive compositionality.

Different training techniques influence the degree of compositionality. In MultiBERTs, compositionality does not steadily increase with training and can decline over time (see Table 3). By contrast, compositionality in knowledge graph embeddings tends to rise steadily during training. Likewise, in SBERT, compositionality increases in earlier layers but declines slightly in later layers, presumably because these upper layers specialize in task-specific representations (Devlin et al., 2018; Tenney et al., 2019a). For graph embeddings, both additive and multiplicative scoring functions perform similarly, and further correlation-based analysis sheds light on the semantic meaning of embedding dimensions.

Our findings address the broader question of whether relations between structured entities can be captured through simple vector operations. Evaluating a range of word, sentence, and graph embeddings showed a common thread of additive compositionality. For sentence embeddings, conceptual components can often be combined additively, indicating that even large, non-linear models (e.g., GPT or Llama) retain interpretable vector structures. Word embeddings effectively encode both morphological and semantic relationships, as illustrated by transformations like *weight + y + ly = weightily.* In knowledge graph embeddings, attributes such as user demographics (age, gender) can be adjusted via vector operations, enabling inference of new relationships in recommender systems.

Overall, these results underscore that embeddings from diverse sources maintain a surprising degree of additive compositional structure, which can be leveraged to enhance interpretability in representation learning. Future work will build on this foundation by exploring debiasing techniques (Bose & Hamilton, 2019) and examining how large language models produce and utilize additive components in tasks like reasoning. Additionally, given the large size of LLaMA embeddings, we plan to explore dimensionality reduction techniques, such as those inspired by the Johnson–Lindenstrauss lemma, to preserve statistical properties while improving numerical stability and computational efficiency. A more detailed investigation at both the layer and attention-head level, as well as an extension to additional graph embedding models, may further illuminate how different architectures preserve and exploit these additive relationships.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2016.

Jacob Andreas. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

Avishek Bose and William Hamilton. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*, pp. 715–724. PMLR, 2019.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Grzegorz Chrupała and Afra Alishahi. Correlating neural and symbolic representations of language. *arXiv preprint arXiv:1905.06401*, 2019.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*, 2018.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv preprint arXiv:1810.04805.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, pp. 134–139, 2016.

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7332–7345, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.595. URL `https://aclanthology.org/2020.emnlp-main.595`.

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

Chenfeng Guo and Dongrui Wu. Canonical correlation analysis (cca) based multi-view learning: An overview. *arXiv preprint arXiv:1907.01693*, 2019.

Zhijin Guo, Zhaozhen Xu, Martha Lewis, and Nello Cristianini. Extract: Explainable transparent control of bias in embeddings. In *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*, 2023. URL `https://aequitas-aod.github.io/aequitas-ecai23.github.io/`.

Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. OpenKE: An open toolkit for knowledge embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 139–144, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2024. URL `https://aclanthology.org/D18-2024`.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Najoung Kim and Tal Linzen. Compositionality as directional consistency in sequential neural networks. In *Workshop on Context and Compositionality in Biological and Artificial Neural Systems*, 2019.

Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, 2020.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pp. 2873–2882. PMLR, 2018.

Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Michael Lepori and R. Thomas McCoy. Picking BERT's brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3637–3651, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.325. URL `https://aclanthology.org/2020.coling-main.325`.

Michael Lepori, Thomas Serre, and Ellie Pavlick. Break it down: Evidence for structural compositionality in neural networks. *Advances in Neural Information Processing Systems*, 36:42623–42660, 2023.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Icml*, 2011.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2021.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8689–8696, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon, and Maximiliano A Wilson. Morpholex: A derivational morphological database for 70,000 english words. *Behavior research methods*, 50:1568–1580, 2018.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, et al. The multiberts: Bert reproductions for robustness analysis. In *10th International Conference on Learning Representations, ICLR 2022*, 2022.

Jose A Seoane, Colin Campbell, Ian NM Day, Juan P Casas, and Tom R Gaunt. Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS computational biology*, 10(10):e1003876, 2014.

Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. Additive compositionality of word vectors. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 387–396, 2019.

John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Vered Shwartz and Ido Dagan. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419, 2019.

Paul Soulos, R. Thomas McCoy, Tal Linzen, and Paul Smolensky. Discovering the compositional structure of vector representations with role learning networks. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 238–254, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.23. URL `https://aclanthology.org/2020.blackboxnlp-1.23`.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*, 2019a.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pp. 2071–2080. PMLR, 2016.

villmow. Github, 2019. URL `https://github.com/villmow/datasets_knowledge_embedding`.

Alexei Vinokourov, Nello Cristianini, and John Shawe-Taylor. Inferring a semantic representation of text via cross-language correlation analysis. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL `https://proceedings.neurips.cc/paper/2002/file/d5e2fbef30a4eb668a203060ec8e5eef-Paper.pdf`.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Similarity analysis of contextual word representation models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4638–4655, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.422. URL `https://aclanthology.org/2020.acl-main.422`.

Zhaozhen Xu, Zhijin Guo, and Nello Cristianini. On compositionality in data embedding. In *Advances in Intelligent Data Analysis XXI: 21st International Symposium, IDA 2023*. Springer, 2023.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

Lang Yu and Allyson Ettinger. Assessing phrasal representation and composition in transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4896–4907, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.397. URL `https://aclanthology.org/2020.emnlp-main.397`.

A Details of Correlation-Based Compositionality Detection Method

Canonical Correlation Analysis (CCA) is used to measure the correlation information between two multivariate random variables (Shawe-Taylor et al., 2004). Just like the univariate correlation coefficient, it is estimated on the basis of two aligned samples of observations.

A matrix of binary-valued attribute embeddings, denoted as $\mathbf{A}$, is essentially a matrix representation where each row corresponds to a specific attribute and each column corresponds to an individual data point (such as a word, image, or user). The entries of the matrix can take only two values, typically 0 or 1, signifying the absence or presence of a particular attribute. For example, in the context of textual data, an attribute might represent whether a word is a noun or not, and the matrix would be populated with 1s (presence) and 0s (absence) accordingly.

On the other hand, a matrix of user embeddings, denoted as $\mathbf{U}$, is a matrix where each row represents an individual user, and each column represents a certain feature or dimension of the embedding space. These embeddings are continuous-valued vectors that capture the movie preference of the users. The values in this matrix are not constrained to binary values and can span a continuous range.

These paired random variables are often different descriptions of the same object, for example genetic and clinical information about a set of patients (Seoane et al., 2014), French and English translations of the same document (Vinokourov et al., 2002), and even two images of the same object from different angles (Guo & Wu, 2019).

In the example of viewers and movies, we use this method to compare two descriptions of users. One matrix is based on demographic information, which are indicated by Boolean vectors. The other matrix is based on their behaviour, which is computed by their movie ratings only.

Assuming we have a vector for an individual user's attribute embedding, denoted as:

$$\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$$

and a corresponding individual user's computed embedding:

$$\mathbf{u} = (u_1, u_2, \ldots, u_m)^T$$

we aim to explore the correlation between these two representations across multiple users. Given a set of q users, we define $\mathbf{A}$ as a $q \times n$ matrix where each row corresponds to the attribute embeddings for a specific user, and $\mathbf{U}$ as a $q \times m$ matrix where each row represents the computed embedding of the same user. Here, n is the number of attributes, and m is the dimensionality of the user embeddings.

Canonical Correlation Analysis (CCA) is then employed to find the projection matrices $\mathbf{W}_A \in \mathbb{R}^{n\times k}$ and $\mathbf{W}_U \in \mathbb{R}^{m\times k}$ that maximise the correlation between the transformed representations of $\mathbf{A}$ and $\mathbf{U}$. Each projection matrix contains k-projection vectors, where k is the number of canonical components that depend on the eigenvalues of the covariance matrix. For each k-th canonical component, the projections of the attribute and user embeddings are given by:

$$\mathbf{A}_k = \mathbf{A}\mathbf{w}_{a_k} \quad \text{and} \quad \mathbf{U}_k = \mathbf{U}\mathbf{w}_{u_k}$$

where $\mathbf{w}_{a_k} \in \mathbb{R}^n$ and $\mathbf{w}_{u_k} \in \mathbb{R}^m$ are the k-th projection vectors from the matrices $\mathbf{W}_A$ and $\mathbf{W}_U$, respectively. $\mathbf{A}_k \in \mathbb{R}^q$ represents the projection of the original attribute matrix $\mathbf{A}$ (size $q \times n$) onto the k-th canonical direction, using the projection vector $\mathbf{w}_{a_k} \in \mathbb{R}^n$. It results in a vector of size $q \times 1$ that contains the transformed values for each user for the k-th component. Similarly, $\mathbf{U}_k \in \mathbb{R}^q$ represents the projection of the original user embedding matrix $\mathbf{U}$ (size $q \times m$) onto the k-th canonical direction, using the projection vector $\mathbf{w}_{u_k} \in \mathbb{R}^m$. It also results in a vector of size $q \times 1$ that contains the transformed values for each user for the k-th component.

The goal of CCA is to maximise the Pearson Correlation Coefficient (PCC) between these transformed representations, i.e., between $\mathbf{A}_k$ and $\mathbf{U}_k$, for each k-th canonical component. The correlation for the k-th canonical component, denoted as ρ_k, is given by the formula:

$$\rho_k = \frac{\sum_{i=1}^{q} \left((\mathbf{A}_k)_i - \mu_{\mathbf{A}_k}\right)\left((\mathbf{U}_k)_i - \mu_{\mathbf{U}_k}\right)}{\sqrt{\sum_{i=1}^{q} \left((\mathbf{A}_k)_i - \mu_{\mathbf{A}_k}\right)^2}\sqrt{\sum_{i=1}^{q} \left((\mathbf{U}_k)_i - \mu_{\mathbf{U}_k}\right)^2}} \tag{1}$$

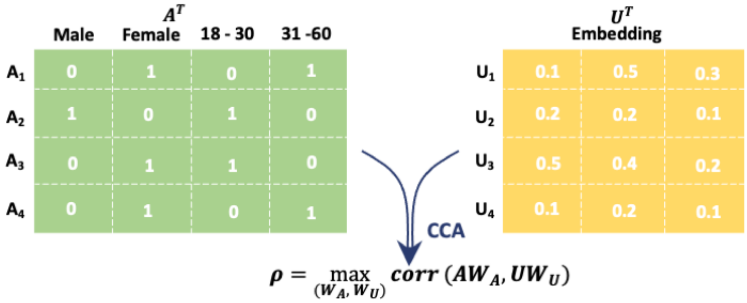
where $\mu_{\mathbf{A}_k}$ and $\mu_{\mathbf{U}_k}$ are the means of the transformed attribute and user embeddings, respectively, and q is the number of users.

In matrix form, we can express this objective as maximising the correlation between the transformed matrices $\mathbf{AW}_A$ and $\mathbf{UW}_U$. This is formalised as:

$$\rho = \max_{\mathbf{W}_A, \mathbf{W}_U} \operatorname{corr}\left(\mathbf{AW}_A, \mathbf{UW}_U\right) \tag{2}$$

where the correlation is maximised across the projection matrices $\mathbf{W}_A$ and $\mathbf{W}_U$, and the result is a set of K-canonical correlations $\rho_1, \rho_2, \ldots, \rho_k$ that describe the relationship between the attribute embeddings and the user embeddings for the entire dataset.

Thus, by computing the canonical correlations ρ_K for each component, we obtain insights into how well the attribute embeddings and computed user embeddings are aligned in terms of their underlying structure.



A^T	Male	Female	18 - 30	31 -60
A_1	0	1	0	1
A_2	1	0	1	0
A_3	0	1	1	0
A_4	0	1	0	1

U^T	Embedding		
U_1	0.1	0.5	0.3
U_2	0.2	0.2	0.1
U_3	0.5	0.4	0.2
U_4	0.1	0.2	0.1

Figure 8: Schematic of Correlation-based compositionality Detection (Guo et al., 2023)

As shown in Figure 8, consider the case where we have 4 users. The attribute matrix $\mathbf{A}$ has dimensions 4×4, representing 4 users and 4 attributes, and the user embedding matrix $\mathbf{U}$ has dimensions 4×3, representing 4 users and 3 embedding dimensions. Thus: - $\mathbf{A} \in \mathbb{R}^{4\times4}$: the attribute matrix where each row represents the 4-dimensional attribute vector for a user. - $\mathbf{U} \in \mathbb{R}^{4\times3}$: the user embedding matrix where each row represents the 3-dimensional embedding vector for a user.

Canonical Correlation Analysis (CCA) is employed to find the projection matrices $\mathbf{W}_A \in \mathbb{R}^{4\times k}$ and $\mathbf{W}_U \in \mathbb{R}^{3\times k}$ that maximize the correlation between the transformed representations of $\mathbf{A}$ and $\mathbf{U}$. Each projection matrix contains k-projection vectors, where k is the number of canonical components that depend on the eigenvalues of the covariance matrix. For each k-th canonical component, the projections of the attribute and user embeddings are given by:

$$\mathbf{A}_k = \mathbf{A}\mathbf{w}_{a_k} \quad \text{and} \quad \mathbf{U}_k = \mathbf{U}\mathbf{w}_{u_k}$$

where $\mathbf{w}_{a_k} \in \mathbb{R}^4$ and $\mathbf{w}_{u_k} \in \mathbb{R}^3$ are the k-th projection vectors from the matrices $\mathbf{W}_A$ and $\mathbf{W}_U$, respectively.

B Leave-one-out experiment

1. **Leave Out**: Exclude the i-th row from $\mathbf{A}$ and $\mathbf{U}$ to obtain $\mathbf{A}_{-i}$ and $\mathbf{U}_{-i}$.
2. **Train**: Solve $\mathbf{A}_{-i}\mathbf{X} = \mathbf{U}_{-i}$ using the pseudo-inverse to obtain $\mathbf{X}$.
3. **Predict**: Estimate the left-out embedding using $\hat{\mathbf{u}}_i = \mathbf{a}_i\mathbf{X}$, where $\mathbf{a}_i$ is the attribute vector of entity i.

4. **Evaluate**: Compare $\hat{\mathbf{u}}_i$ with the actual embedding $\mathbf{u}_i$ using the following metrics:
    (a) *L2 Loss*: $L_2 = \|\hat{\mathbf{u}}_i - \mathbf{u}_i\|^2$, measuring the reconstruction error.
    (b) *Embedding Prediction (Cosine Similarity)*: $\cos(\theta) = \dfrac{\hat{\mathbf{u}}_i \cdot \mathbf{u}_i}{\|\hat{\mathbf{u}}_i\| \, \|\mathbf{u}_i\|}$, assessing the alignment between the predicted and actual embeddings.
    (c) *Identity Prediction (Retrieval Accuracy)*: Determines if $\hat{\mathbf{u}}_i$ correctly identifies entity i by checking if $\mathbf{u}_i$ is the nearest neighbor to $\hat{\mathbf{u}}_i$ among all embeddings.

C Knowledge Graph Embedding

Knowledge Graph Embedding A *graph* $G = (V, E)$ consists of a set of vertices V with edges E between pairs of vertices. In a *knowledge graph*, the vertices V represent entities in the real world, and the edges E encode that some relation holds between a pair of vertices. As a running example, we consider the case where the vertices V are a set of viewers and films, and the edges E encode the fact that a viewer has rated a film.

Knowledge Graphs represent information in terms of entities (or nodes) and the relationships (or edges) between them. The specific relation r that exists between two entities is depicted as a directed edge, and this connection is represented by a triple (h, r, t). In this structure, we distinguish between the two nodes involved: the *head* (h) and the *tail* (t), represented by vectors $\mathbf{h}$ and $\mathbf{t}$ respectively. Such a triple is termed a *fact*, denoted by f:

$$f = (h, r, t)$$

Embedding Function. A knowledge graph $G = (V, E)$ can be embedded by assigning each node $v \in V$ a vector $\mathbf{x}_v \in \mathbb{R}^d$. This embedding function, $\Phi_{KG} : V \to \mathbb{R}^d$, maps nodes into a continuous space where their relationships are captured by a scoring or distance function. For instance, one could define a threshold θ such that an edge $(v_i, v_j) \in E$ exists if and only if $D(\mathbf{x}_{v_i}, \mathbf{x}_{v_j}) < \theta$. Conversely, given a set of embedded points, links between nodes can be recovered by applying a learned scoring function $S\big(\mathbf{x}_{v_i}, \mathbf{x}_{v_j}\big)$.

C.1 Multiplicative Scoring

Nickel et al. (2011) introduced a tensor-factorization approach for relational learning, treating each frontal slice of a three-dimensional tensor as a co-occurrence matrix for a given relation. In this model, a triple (h, r, t) is scored using embeddings $\mathbf{h}, \mathbf{R}, \mathbf{t}$, where $\mathbf{h}, \mathbf{t} \in \mathbb{R}^d$ represent head and tail entities, and $\mathbf{R} \in \mathbb{R}^{d \times d}$ represents the relation:

$$S(h, r, t) \; = \; \mathbf{h}^T \, \mathbf{R} \, \mathbf{t}. \tag{3}$$

Various specializations exist, such as DistMult (Yang et al., 2014), which restricts $\mathbf{R}$ to a diagonal matrix (reducing overfitting), and ComplEx (Trouillon et al., 2016), which employs complex-valued embeddings for asymmetric relations. In this work, we adopt DistMult due to its simplicity and scalability, particularly its suitability for large knowledge graphs.

C.2 Additive Scoring

TransE (Bordes et al., 2013) introduces a translation-based perspective, where each relation is a vector that shifts the embedding of the head entity to the tail entity. A triple (h, r, t) is scored by:

$$S(h, r, t) \; = \; \|\mathbf{h} + \mathbf{r} \; - \; \mathbf{t}\|, \quad \mathbf{h}, \mathbf{r}, \mathbf{t} \; \in \; \mathbb{R}^d. \tag{4}$$

For example, *King* + *FemaleOf* ≈ *Queen*. This translation idea captures relational semantics by minimizing the distance between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$.

Rating Prediction In alignment with (Berg et al., 2017), we establish a function P that, given a triple of embeddings $(\mathbf{h}, \mathbf{R}, \mathbf{t})$, calculates the probability of the relation against all potential alternatives.

$$P\left(\mathbf{h}, \mathbf{R}, \mathbf{t}\right) = \text{SoftArgmax}(S(f)) = \frac{e^{S(f)}}{e^{S(f)} + \sum_{r' \neq r \in \mathscr{R}} e^{S(f')}} \tag{5}$$

In the above formula, $f = (h, r, t)$ denotes a true triple, and $f' = (h, r', t)$ denotes a corrupted triple, that is a randomly generated one, that we use as a proxy for a negative example (a pair of nodes that are not connected).

Assigning numerical values to relations r, the predicted relation is then just the expected value prediction $= \sum_{r \in \mathscr{R}} r P\left(\mathbf{h}, \mathbf{R}, \mathbf{t}\right)$ In our application of viewers and movies, the set of relations $\mathscr{R}$ could be the possible ratings that a user can give a movie. The predicted rating is then the expected value of the ratings, given the probability distribution produced by the scoring function. $S(f)$ refers to the scoring function in Yang et al. (2014).

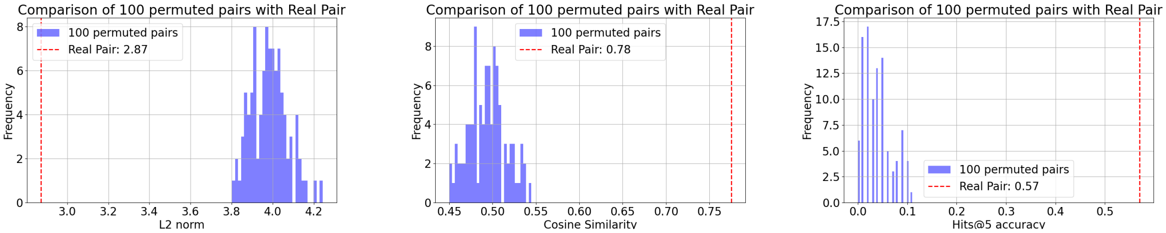
To learn a graph embedding, we follow the setting of Bose & Hamilton (2019) as follows,

$$L = -\sum_{f \in \mathscr{F}} \log \frac{e^{S(f)}}{e^{S(f)} + \sum_{f' \in \mathscr{F}'} e^{S(f')}} \tag{6}$$

This loss function maximizes the probabilities of true triples (f) and minimizes the probability of triples with corrupted triples: (f').

Evaluation Metrics We use 4 metrics to evaluate our performance on the link prediction task. These are root mean square error (RMSE, $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}$, where $\hat{y}_i$ is our predicted relation and y_i is the true relation), Hits@K - the probability that our target value is in the top K predictions, mean rank (MR) - the average ranking of each prediction, and mean reciprocal rank (MRR) to evaluate our performance on the link prediction task. These are standard metrics in the knowledge graph embedding community.

D Additive Compositionality by Model



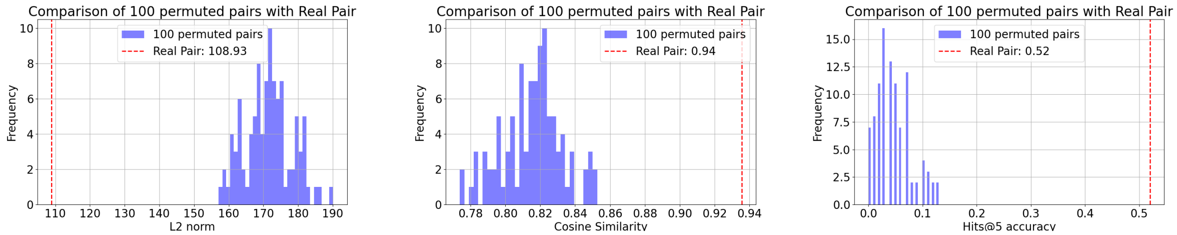
(a) Linear System Loss (b) Cosine Similarity (c) Retrieval Accuracy@5

Figure 9: Test statistics for GPT embedding decomposition. Dashed line is the average performance of $\hat{\mathbf{U}}$ learned from the user embedding. Bars are the distribution of the results from 100 random permutations.

E Compositionality across Layers and Training Stages

E.1 Comparison of Different Layers

Comparison Metrics To fairly compare different layers, we cannot rely solely on raw cosine similarities or retrieval accuracies due to variations in scales and distributions. Instead, we use normalized metrics for comparability. The **Normalized Cosine Similarity** computes the difference between the mean real similarity and the mean permuted similarity, normalized by the maximum possible difference,



(a) Linear System Loss (b) Cosine Similarity (c) Retrieval Accuracy@5

Figure 10: Test statistics for Llama embedding decomposition. Dashed line is the average performance of $\hat{\mathbf{U}}$ learned from the user embedding. Bars are the distribution of the results from 100 random permutations.

(1 − Mean Permuted Similarity). The **Absolute Difference** is a simple measure of the difference between mean real and permuted similarities. Lastly, the **Relative Difference (Percentage Improvement)** expresses this difference as a percentage of the permuted similarity, indicating proportional improvement. These metrics enable robust and fair comparisons across models.

Table 5: Additive Compositionality Metrics Across SBERT Layers

Layer	Mean Sim (Real)	Mean Sim (Permuted)	Norm. Cosine Sim	Hits@5 Acc (Real)	Hits@5 Acc (Permuted)	Norm. Retrieval Acc
0	0.8889	0.7808	0.4930	0.59	0.0397	0.5731
1	0.9366	0.8671	0.5228	0.57	0.0406	0.5518
2	0.9397	0.8576	0.5767	0.61	0.0390	0.5942
3	0.9403	0.8330	0.6424	0.64	0.0412	0.6245
4	0.9408	0.8298	0.6523	0.66	0.0400	0.6458
5	0.9409	0.8273	0.6577	0.62	0.0417	0.6035
6	0.7761	0.4865	0.5640	0.59	0.0405	0.5727

Table 6: Additive Compositionality Metrics at Different Training Steps of BERT

Model	Training Steps	Mean Sim (Real)	Mean Sim (Permuted)	Norm. Cosine Sim	Hits@5 Acc (Real)	Permuted Acc	Norm. Retrieval Acc
cls_0k	0	0.9884	0.9882	0.0163	0.44	0.0418	0.4156
cls_20k	20,000	0.8787	0.7722	0.4676	0.55	0.0407	0.5309
cls_40k	40,000	0.8773	0.7724	0.4607	0.48	0.0405	0.4581
cls_100k	100,000	0.9201	0.8323	0.5236	0.55	0.0417	0.5304
cls_1000k	1,000,000	0.9545	0.9149	0.4655	0.44	0.0408	0.4162
cls_2000k	2,000,000	0.9538	0.9094	0.4896	0.48	0.0415	0.4575

F Decomposing Sentence Embedding into Subject, Verb and Object

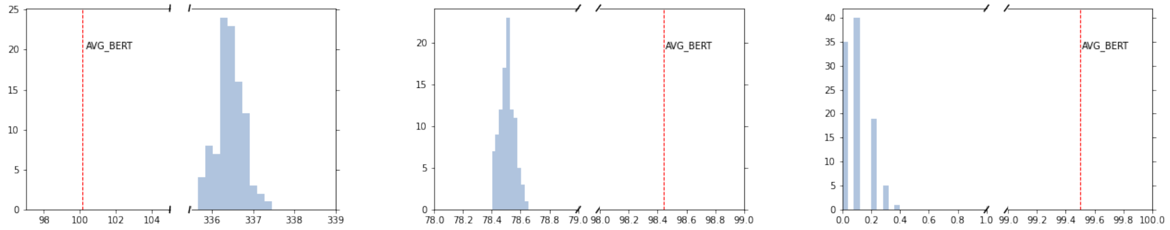
Xu et al. (2023) show that BERT [CLS] tokens can be decomposed into a sum of individual word representations. We recap their results here and report on extensions to their work.

F.0.1 Results

Figure 11 illustrates the performance of decomposing BERT sentence embedding. These results show that the BERT sentence embedding can be decomposed into three separate components: subject, verb, and object. Those components can then be used to predict the embedding of a new sentence.

Furthermore, $\hat{\mathbf{U}}$ achieves a 99.5% success rate in retrieving the correct BERT embedding, whereas the best retrieval accuracy using randomized attribute/embedding pairings does not exceed 0.4%.

While these results are encouraging, it is the case that the other tokens in the sentence are created in the same way as the [CLS] token used for the sentence embedding. We carry out the same experiment across 30 random seeds using each of these other tokens as the representation of the sentence. This forms a more challenging baseline than a random permutation of embedding pairings. In this experiment, we replace some multi-token words with single token words. Specifically, we change “hamster” to “bear”, “hedgehog” to “fox”, “bookshelf” to “book”.



(a) Linear System Loss (b) Cosine Similarity (c) Retrieval Accuracy@1

Figure 11: The test statistics for sentence embedding decomposition. AVG_BERT is the average performance of $\hat{\mathbf{B}}$ learned from the BERT embedding. The bars are the distribution of the results from random permutations that run for 100 times (Xu et al., 2023).

Table 7 shows the performance of the linear system for each token level. We used three metrics to assess compositionality: L2 loss, cosine similarity, and retrieval accuracy.

We see that across all metrics, there are no substantial differences in whether the candidate sentence embedding can be decomposed into subject, verb, and object. The similarity of each token's embedding in the last layer of the transformer suggests a lack of distinct information among them, a finding that is supported by Park & Kim (2021). However, under the key performance metric of retrieval accuracy, the [CLS] token embedding does perform best. We performed a one-sided t-test and found that the retrieval accuracy of the embeddings produced from the [CLS] token is significantly higher ($p < 10^{-4}$) than the retrieval accuracies of the other tokens.

Table 7: Compositionality for each word in the sentence. Note that in this experiment, we replace some multi-token words into single token words. Specifically, we change "hamster" to "bear", "hedgehog" to "fox", "bookshelf" to "book".

Metric	**Values**		
	L2 Loss	**Cosine Similarity**	**Retrieval Accuracy**
CLS	103.91	0.983	0.995
Subject	96.46	0.985	0.988
Verb	108.58	0.980	0.993
Object	102.55	0.983	0.991
The first 'The'	101.01	-	-
The last '.'	97.13	0.984	0.992
Random Baseline of CLS	343	0.77	0.01

Compositionality across layers of SBERT We further investigate the differences between token embeddings through the layers of SBERT. We repeat the same experiment for each token and for each layer of the model. Table 10 reports the metrics for the [CLS] token, and Figures 12, 13, and 14 show the L2 loss, cosine similarity and retrieval accuracy across the 12 layers of SBERT. More detailed results corresponding to these figures are shown in Table 8 and Table 9.

We see that through early layers of SBERT, up to layer 9, the [CLS] token is more amenable to decomposition into component word embeddings than the other tokens are, with lower loss, higher cosine similarity, and higher retrieval accuracy than the other token embeddings. We further see that in the earlier layers of the model, the [CLS] token exhibits more additive compositionality than in later layers. This indicates that as the sentence is processed through the layers, more contextual information is being added. However, what is interesting is that the amount of contextual information being added is still low, and much of the embedding can be accounted for additively. Considering cosine similarity (Figure 13), we see that the cosine similarity

Table 8: Compositionality for Different Words in a Sentence (layer 1 to layer 6)

Metric	**Values**		
	L2 Loss	**Cosine Similarity**	**Retrieval Accuracy**
		Initial Embedding	
		Layer 1	
CLS	1.05	1.0	1.0
Subject	47.13	0.997	0.985
Verb	55.31	0.996	0.97
Object	46.18	0.997	0.956
		Layer 2	
CLS	1.99	1.0	1.0
Subject	69.10	0.995	0.918
Verb	66.05	0.995	0.942
Object	64.00	0.996	0.923
		Layer 3	
CLS	3.79	1.0	1.0
Subject	69.11	0.995	0.871
Verb	68.30	0.994	0.832
Object	65.73	0.995	0.843
		Layer 4	
CLS	10.72	1.0	1.0
Subject	67.53	0.995	0.84
Verb	77.01	0.993	0.798
Object	66.53	0.995	0.84
		Layer 5	
CLS	18.64	0.999	0.996
Subject	73.40	0.994	0.84
Verb	82.05	0.992	0.804
Object	71.46	0.995	0.858
		Layer 6	
CLS	21.88	0.999	0.985
Subject	78.51	0.993	0.85
Verb	85.13	0.992	0.72
Object	73.38	0.994	0.826

Table 9: Compositionality for Different Words in a Sentence (layer 7 to layer 12)

Metric	**Values**		
	L2 Loss	**Cosine Similarity**	**Retrieval Accuracy**
		Layer 7	
CLS	34.64	0.998	0.981
Subject	72.49	0.994	0.826
Verb	84.36	0.991	0.717
Object	70.17	0.995	0.835
		Layer 8	
CLS	25.48	0.999	0.959
Subject	62.49	0.994	0.821
Verb	74.51	0.992	0.685
Object	62.28	0.995	0.85
		Layer 9	
CLS	30.22	0.997	0.949
Subject	56.14	0.994	0.865
Verb	68.16	0.991	0.793
Object	54.89	0.994	0.887
		Layer 10	
CLS	54.01	0.993	0.98
Subject	66.86	0.992	0.951
Verb	86.81	0.987	0.922
Object	64.43	0.993	0.974
		Layer 11	
CLS	76.66	0.988	0.996
Subject	72.46	0.991	0.987
Verb	96.29	0.985	0.976
Object	79.55	0.990	0.992
		Layer 12	
CLS	103.91	0.983	0.995
Subject	96.46	0.985	0.988
Verb	108.58	0.980	0.993
Object	102.55	0.983	0.991

of the [CLS] embedding with its reconstruction drops from 1 in very early layers to around 0.98, indicating that a large proportion of the composition in the sentence can be interpreted additively.

Each transformer layer in SBERT consists of two main components: a multi-head self-attention mechanism and a position-wise feed-forward network. This contributes to refining the representations, making them richer and more context-aware. We are interested in measuring the compositionality in these represenatations of different layers. Figure 12, 13, 14show the L2 loss, cosine similarity and retrieval accuracy across 12 layers of SBERT Embedding.

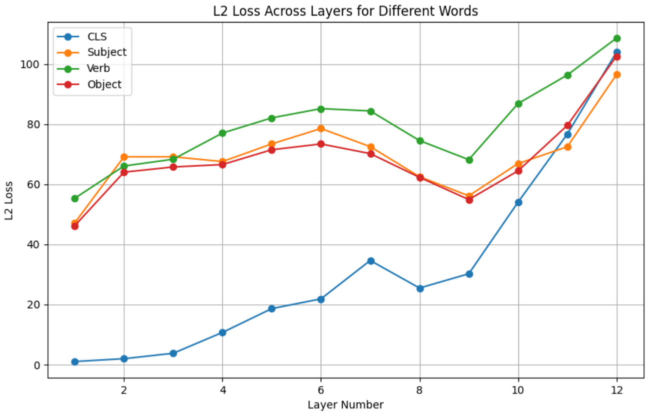


Figure 12: L2 Loss Across Layers in SBERT Embedding

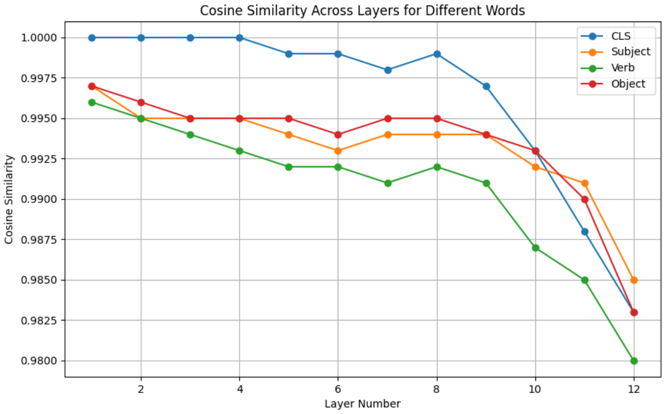


Figure 13: Cosine Similarity Across Layers in SBERT Embedding

Compositionality across BERT training stages. We also looked into how compositionality is captured during different training stages of BERT. We used the MultiBERTs Sellam et al. (2022) to get intermediate checkpoints captured during pre-training steps. Results are shown in Table 11 and Figure 15.

We see that at the beginning of training, the cosine similarity between the [CLS] embedding and the reconstructed embedding is very high, with perfect retrieval accuracy. We conjecture that this is due to the initialization of the BERT model. Sellam et al. (2022) use a GELU activation function and initialize the model with weights drawn from a truncated normal distribution with mean 0 and standard deviation 0.02. Close to 0, the GELU activation function is approximately linear. This means that the [CLS] token will naturally decompose into a weighted sum of component vectors.

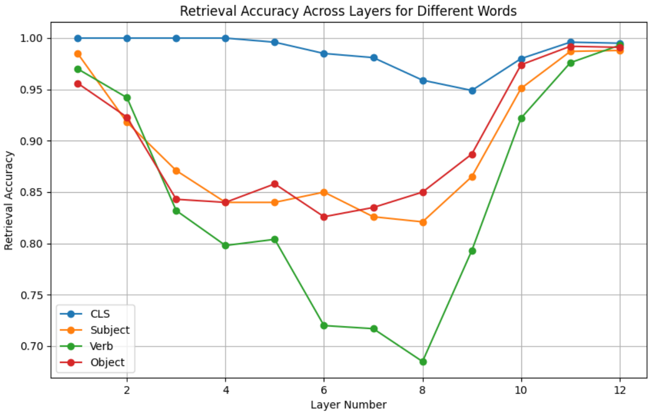


Figure 14: Retrieval Accuracy Across Layers in SBERT Embedding

Table 10: Compositionality for [CLS] Token in a Sentence (layer 1 to layer 12)

Metric	**Values**		
	L2 Loss	**Cosine Similarity**	**Retrieval Accuracy**
Layer 1	1.05	1.0	1.0
Layer 2	1.99	1.0	1.0
Layer 3	3.79	1.0	1.0
Layer 4	10.72	1.0	1.0
Layer 5	18.64	0.999	0.996
Layer 6	21.88	0.999	0.985
Layer 7	34.64	0.998	0.981
Layer 8	25.48	0.999	0.959
Layer 9	30.22	0.997	0.949
Layer 10	54.01	0.993	0.98
Layer 11	76.66	0.988	0.996
Layer 12	103.91	0.983	0.995

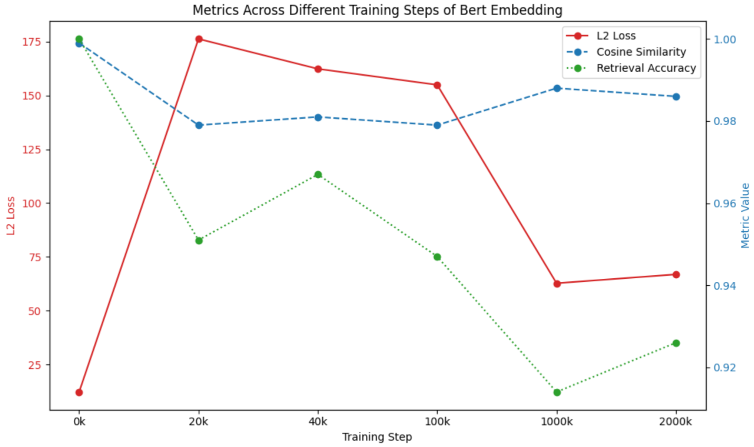


Figure 15: Metrics Across Different Training Steps of Bert Embedding

Table 11: Compositionality in Different Training Steps of Bert Embedding

Training Step	**Values**		
	L2 Loss	**Cosine Similarity**	**Retrieval Accuracy**
CLS_0k	12.23	0.999	1.0
CLS_20k	176.24	0.979	0.951
CLS_40k	162.33	0.981	0.967
CLS_100k	154.90	0.979	0.947
CLS_1000k	62.74	0.988	0.914
CLS_2000k	66.85	0.986	0.926
Random Baseline of CLS_2000k	141	0.93	0.01

As training progresses, we see that cosine similarity between the [CLS] token and its reconstruction decreases, meaning that more contextual information is added. However, what is interesting is that this again does not decrease by a large amount - the main component of the composition in these sentences is still additive.

G Word Embeddings

G.1 WordNet Embedding

We want to ensure our WordNet embedding can contain the semantic relation in it. Therefore, we train the embedding with the task of predicting the tail entity given a head entity and relation. For example, we might want to predict the hypernym of cat:

$$< \textbf{cat}, hypernym, \textbf{?} >$$

Mapping Freebase ID to text WordNet is constructed with Freebase ID only, an example triple could be <00260881, hypernym, 00260622>. We follow villmow (2019) to preprocess the data and map each entity with the text with a real meaning.

The above triple can then be processed with the real semantic meaning: <land reform, hypernym, reform>. The word2vec word embedding is pretrained from a google news corpus. We train the WordNet Embedding in the following way:

1. We split our dataset to use 90% for training, 10% for testing.

2. Triples of $(head, relation, tail)$ are encoded as relational triples (h, r, t).
3. We randomly initialize embeddings for each h_i, r_j, t_k, use the scoring function in Equation 4 and minimize the loss by Margin Loss.
4. We sampled 20 corrupted entities. Learning rate is set at 0.05 and training epoch at 300.

Results can be found in the Table 12, which shows that our WordNet embeddings do contain semantic information.

Table 12: Link prediction performance for WordNet

	Hits@1	Hits@3	Hits@10	MRR
WordNet	0.39	0.41	0.43	0.40

H List of words of experiments: Decomposing word2vec Embedding by Additive Compositionality Detection

allegorically, whimsicality, whimsically, voyeuristically, weightier, weightiest, weightily, weightiness, weightlessly, veritably, visualizations, tyrannically, traitorously, transcendentally, transitionally, tangentially, temperamentally, surgically, structurally, studiously, studiousness, stylistically, spiritualistic, slipperiest, slipperiness, serviceability, serviceably, sectionalism, sentimentalism, sentimentalist, sentimentalized, sentimentalizes, sentimentalizing, sentimentally, serialization, serializations, satanically, reverentially, ritualistically, regularization, quizzically, rapturously, puritanically, probationers, probationer, psychiatrically, preferentially, practicably, practicalities, pleasurably, polarizations, phenomenally, personalizations, pessimistically, pathetically, occupationally, optionally, oratorically, nationalizations, nautically, neutralizer, neutralizers, mysteriously, mysteriousness, narcissistically, moralistically, melodiously, melodiousness, memorializes, memorializing, metrication, materialistically, mechanistically, mechanizations, longitudinally, lexically, liberalization, liquidator, liquidators, journalistic, inferentially, injuriously, hysterically, idealistically, heretically, futuristically, fractionally, fluoridation, fictionalized, fictionalizes, fictionalizing, figuratively, farcically, fatalistic, environmentalists, environmentally, episodically, equitably, emotionlessly, ecclesiastically, editorialized, editorializes, editorializing, editorially, educationalist, educationalists, educationally, egotistical, egotistically, dictatorships, differentiations, derivatively, developmentally, deviationist, deviationists, decoratively, deferentially, definitively, demagogically, demonically, decimalization, cumulatively, conversationalists, conversationally, confidentialities, conspiratorially, collectivization, colonialists, classically, chauvinistically, censoriousness, certifications, capitalizations, breathalyser, breathalysers, brutalization, antagonistically, apocalyptically, weightlessness, westernization, victoriously, visualization, vocalization, urbanization, Unitarian, Unitarians, transcendentalism, transcendentalist, transcendentalists, theatrically, theoretically, technicalities, technicality, technically, speculatively, socialistic, socialization, sophisticate, sophisticated, sophisticates, significantly, sensational, sentimentalists, sentimentality, sentimentalize, scientifically, satirically, rotationally, residentially, relativistic, realistically, prudentially, pressurization, probabilistic, probationary, professionalism, professionally, popularization, potentialities, potentiality, potentially, practicability, practicality, practically, polarization, phosphorescence, phosphorescent, physically, physicalness, periodically, personalization, particularistic, paternalistic, operational, oratorical, organizational, normalization, numerically, negatively, negativism, neutralization, nationalistic, nationalization, naturalistic, naturalization, mechanically, memorialize, memorialized, metrically, maturational, localization, linguistically, liquidation, liquidations, juridical, justifiably, industrialization, imperialistic, incidentally, identically, imaginatively, historically, harmoniously, graphically, generalization, generalizations, fictionalize, existentialism, existentialist, existentialists, evangelicalism, environmentalism, environmentalist, equalization, equalizers, equatorial, electrically, electronically, emotionalism, emotionality, emotionally, economically, editorialist, editorialize, directionality, directionally, dictatorial, dictatorship, differentiation, conversationalist, confidentiality, confidentially, colonialism, colonialist, commercialization, communicational, civilizational, centralization, certification, chemically, capitalistic, capitalization, catastrophically, categorically, behaviorally, authentication, authentications, authenticator, artistically, architecturally, anatomically, Anglicanism, alternatively, altruistically, adventurously, acoustically, activation, additionally

I Movie-Lens Training Details

This experiment was conducted on the MovieLens 1M dataset (Harper & Konstan, 2015) which consists of a large set of movies and users, and a set of movie ratings for each individual user. It is widely used to create and test recommender systems. Typically, the goal of a recommender system is to predict the rating of an unrated movie for a given user, based on the rest of the data. The dataset contains 6040 users and approximately 3900 movies. Each user-movie rating can take values in 1 to 5. There are 1 million triples (out of a possible $6040 \times 3900 = 23.6m$), so that the vast majority of user-movie pairs are not rated.

Users and movies each have additional attributes attached. For example, users have demographic information such as gender, age, or occupation. Whilst this information is typically used to improve the accuracy of recommendations, we use it to test whether the embedding of a user correlates to private attributes, such as gender or age. We compute our graph embedding based only on ratings, leaving user attributes out. Experiments for training knowledge graph embeddings are implemented with the OpenKE (Han et al., 2018) toolkit. We train our model on GeForce GTX TITAN X.

We embed the knowledge graph in the following way:

1. We split our dataset to use 90% for training, 10% for testing.
2. Triples of $(user, rating, movie)$ are encoded as relational triples (h, r, t).
3. We randomly initialize embeddings for each h_i, r_j, t_k and train embeddings to minimize the loss in equation 6 [7].
4. We sampled 10 corrupted entities and 4 corrupted relations per true triple. Learning rate is set at 0.01 and training epoch at 300.

We verify the quality of the embeddings by carrying out a link prediction task on the remaining 10% test set. We achieved a RMSE score of 0.88, Hits@1 score of 0.46 and Hits@3 as 0.92, MRR as 0.68 and MR as 1.89.

J Analysis of the First Canonical Component of CCA for Knowledge Graph Embedding

---

[7] We add a negative sign for the additive scoring function, since we want to maximize the probability of the true triple, which aligns the setting of this loss function

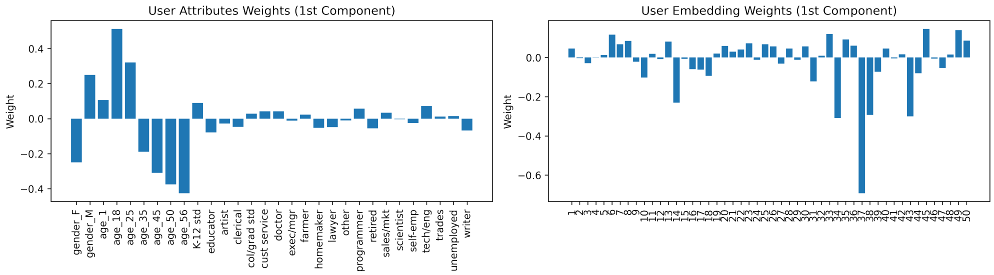


Figure 16: Positive weights in the first canonical component of CCA show the association between user attributes (e.g., *male* and *Age: 18*) and user embedding dimensions (e.g., *44* and *49*).