# Quasi-symbolic Semantic Geometry over Transformer-based Variational AutoEncoder

**Yingji Zhang**[1†], **Danilo S. Carvalho**[1,3], **André Freitas**[1,2,3]
[1] Department of Computer Science, University of Manchester, United Kingdom
[2] Idiap Research Institute, Switzerland
[3] National Biomarker Centre, CRUK-MI, Univ. of Manchester, United Kingdom
{firstname.lastname}@[postgrad.][†]manchester.ac.uk

## Abstract

Formal/symbolic semantics can provide canonical, rigid controllability and interpretability to sentence representations due to their *localisation* or *composition* property. How can we deliver such property to the current distributional sentence representations to better control and interpret the generation of language models (LMs)? In this work, we theoretically frame the sentence semantics as the composition of *semantic role - word content* features and propose the formal semantic geometrical framework. To inject such geometry into Transformer-based LMs (i.e. GPT2), we deploy a supervised Transformer-based Variational AutoEncoder, where the sentence generation can be manipulated and explained over low-dimensional latent Gaussian space. In addition, we propose a new probing algorithm to guide the movement of sentence vectors over such geometry. Experimental results reveal that the formal semantic geometry can potentially deliver better control and interpretation to sentence generation.

## 1 Introduction

Language Models (LMs) have provided a flexible scaling-up foundation for addressing a diverse spectrum of tasks (Touvron et al., 2023). Nonetheless, the question remains: can we develop language representations/models that offer more granular levels of control and interpretation from the perspective of "formal/structural" semantics? Addressing this question will enable us to enhance the controllability, interpretability, and safety of LMs.

Formal semantics, which provides a canonical, granular, and rigid representation, have been investigated for thousands of years with well established theoretical frameworks, such as Montague Semantics (Dowty et al., 2012), Davidsonian Semantics (Davidson, 1967), Semantic Role Labelling (SRL, Palmer et al. (2010)), and Argument Structure Theory (AST, Jackendoff (1992)). One typical characteristic of such formal semantics is the *localisation* or *composition* property. For example, in the sentence: *animals require oxygen for survival*, the words are functionally combined into sentence semantics: $\lambda x(\text{animals}(x) \rightarrow \text{require}(x, \text{oxygen}))$ where $x$ is the variable of any entity within a logical structure. In this case, we can localise the sentence semantics by replacing $x$ with *birds*, etc. This localised process indicates the interpretation in Cognitive Science (Lees, 1957; Smolensky, 2006). However, such localisation is precisely what current distributional semantics lack, thereby limiting their controllability and interpretability.

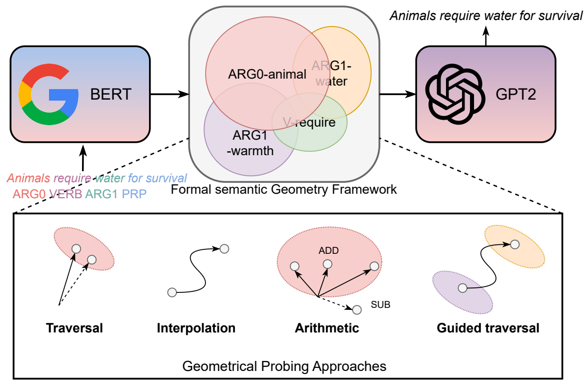



Figure 1: Overview: latent sentence semantics can be decomposed into *semantic role- word content* features.

Disentanglement (Bengio, 2013), which refers to the feature-dimension alignment, can potentially provide such localisation, which has been widely investigated to localise image features, such as *nose* in facial images (Esser et al., 2020; Jeon et al., 2019; Liu et al., 2021). In Transformers (Vaswani et al., 2017), however, token embeddings, residual stream, and attention have the *polysemanticity* phenomenon (Elhage et al., 2022), meaning that multiple dimensions contribute to a feature. Although some prior studies explored the possibility of language disentanglement, most are focused on coarse-grained/task-specific semantic features, such as sentiment, within the context of style-transfer tasks



*Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 12–29
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

(John et al., 2019; Bao et al., 2019; Hu and Li, 2021; Vasilakes et al., 2022; Gu et al., 2022; Liu et al., 2023a; Gu et al., 2023).

In this work, we focus on the localisation of *general* semantic features of sentences over distributional space to shorten the gap between deep latent semantics and formal linguistic representations (Gildea and Jurafsky, 2000; Banarescu et al., 2013; Mitchell, 2023), integrating the flexibility of distributional-neural models with the properties of linguistically grounded representations, facilitating both interpretability and generative control from the perspective of formal semantics. We specifically choose the conceptual dense explanatory sentences from WorldTree (Jansen et al., 2018) due to their clear formal semantic representation designed in the explanatory, cognitive reasoning task.

In the NLP domain, Variational AutoEncoders (VAEs, Kingma and Welling (2013)) have been recognized as a prominent foundation for investigating generation control and interpretation through the observable low-dimensional smooth and regular latent spaces (e.g., std Gaussian space) (John et al., 2019; Li et al., 2022b; Bao et al., 2019; Mercatali and Freitas, 2021; Felhi et al., 2022; Vasilakes et al., 2022). Therefore, we probe the localisation property of formal semantics over latent sentence spaces under VAE architecture. Specifically:

**(1)** We first propose a geometrical framework to present the formal semantic features of sentences as *semantic role - word content* pairs (denoted as role-content) from the perspective of AST (Jackendoff, 1992) within the compositional distributional model (Clark et al., 2008). Subsequently, **(2)** we introduce a supervised approach for learning the role-content features of explanatory sentences in latent spaces. **(3)** Additionally, we contribute to a method to control sentence generation by navigating the sentence vectors across different role-content features within our geometric framework. **(4)** Our findings reveal that the role-content features are encoded as a convex cone in the latent sentence space (Figure 1). This semantic geometry facilitates the localisation of sentence generation by enabling the manipulation of sentence vectors through traversal and arithmetic operations within the latent space.

## 2 Related work

**Formal-distributional semantics.** Integrating distributional semantics with formal / symbolic semantics is challenging due to the difficulty of optimisation over discrete space (van Krieken et al., 2023). In the Reasoning domain, existing approaches usually perform symbolic behaviour via explicitly symbolic representation injection, including graph (Khashabi et al., 2018; Khot et al., 2017; Jansen et al., 2017; Thayaparan et al., 2021), linear programming (Valentino et al., 2022b; Thayaparan et al., 2024), adopting iterative methods, using sparse or dense encoding mechanisms (Valentino et al., 2020; Lin et al., 2020; Valentino et al., 2022a; Bostrom et al., 2021), or synthetic natural language expression (Clark et al., 2020; Yanaka et al., 2021; Fu and Frank, 2024), among others. Comparatively, we explore the formal semantic property over distributional semantics via latent sentence geometry, which can potentially deliver better interpretation and control to current LMs.

**Language geometry.** There is a line of work that studies the geometry of word and sentence representations (Arora et al., 2016; Mimno and Thompson, 2017; Ethayarajh, 2019; Reif et al., 2019; Li et al., 2020a; Chang et al., 2022; Jiang et al., 2024a). E.g., $king - man + woman = queen$, which the word vectors can be manipulated with geometric algebra. This phenomenon indicates the linear subspaces in language representations, similar features are encoded as a close direction in latent space, which has been widely explored ranging from word (Mikolov et al., 2013a) to sentences (Ushio et al., 2021), Transformer-based LMs (Merullo et al., 2023; Hernandez et al., 2023), and multi-modal models (Trager et al., 2023; Huh et al., 2024). Under the linear subspace hypotheses, a significant work explored the interpretability (Li et al., 2022a; Geva et al., 2022; Nanda et al., 2023) and controllability (Trager et al., 2023; Merullo et al., 2023; Turner et al., 2023) of neural networks. In this work, we emphasise the formal semantic geometry for bridging the distributional and formal semantics, which is currently under-explored.

**Language disentanglement.** Disentanglement, refers to separating features along dimensions (Bengio, 2013), leading to clear geometric and linear representations. In the NLP domain, prior studies explored the disentanglement between specific linguistic perspectives, such as sentiment-content (John et al., 2019), semantic-syntax (Bao et al., 2019), and negation-uncertainty (Vasilakes et al., 2022), or syntactic-level disentanglement (Mercatali and Freitas, 2021; Felhi et al., 2022). However, those approaches focused on disentangling coarse-grained/task-specific semantic features. In this work, we contribute to a new lens on the disentanglement (separation) of "general" sentence features from the perspective of formal semantics.

## 3 Formal Semantic Geometry

In this section, we first define the sentence semantic features as *semantic role - word content* from the perspective of formal semantics. Then, we link the semantic features with distributional vector spaces in which each *semantic role - word content* is encoded as a convex cone, as shown in Figure 1.

**Formal semantic features.** For formal / structural semantics, *Argument Structure Theory (AST)* (Jackendoff, 1992; Levin, 1993; Rappaport Hovav and Levin, 2008) provides a model for representing sentence structure and meaning of sentences in terms of the interface between the their syntactic structure and the associated semantic roles of the arguments within those sentences. It delineates how verbs define the organisation of their associated arguments and the reflection of this organisation in a sentence's syntactic realisation. AST abstracts sentences as predicate-argument structures, where the predicate $p$ (associated with the verb) has a set of associated arguments $arg_i$, where each argument has an associated positional component $i$ and a thematic/semantic roles $r_i$, the latter categorising the semantic functions of arguments in relation to the verb (e.g. agent, patient, theme, instrument). In the context of this work, the AST predicate-argument representation is associated with a lexical-semantic representation of the content $c_i$ of the term $t_i$.

In this work, we simplify and particularise the relationship between the argument structure and the distributional lexical semantic representation as a *role-content* relation, where the structural syntactic/semantic relationship is defined by its shallow semantics, i.e. as the composition of the content of the terms, their position in the predicate-argument (PArg) structure ($arg_i$) and their semantic roles (SRs) ($r_i$: $pred$, $arg$), as described below:

$$\underbrace{animals}_{ARG0}\ \underbrace{require}_{PRED}\ \underbrace{oxygen}_{ARG1}\ \underbrace{for\ survival}_{ARGM-PRP}$$

Therefore, we define the semantics of sentences, $sem(s)$, as the compositions between *role-content*, which can be described as follows:

$$sem(s) = \underbrace{t_1(c_1, r_1)}_{i.e., ARG0-animals} \oplus \cdots \oplus \underbrace{t_i(c_i, r_i)}_{PRP-survival}$$

Where $t_i(c_i, r_i) = c_i \otimes r_i$ represents the semantics of term $t_i$ with content $c_i$ (i.e., *animals*) and SRL $r_i$ (i.e., *ARG0*) in context $s$. $\otimes$: connects the meanings of words with their roles, using the compositional-distributional semantics notation of (Smolensky and Legendre, 2006; Clark and Pulman, 2007; Clark et al., 2008). $\oplus$: connects the lexical semantics (word content + structural role) to form the sentence semantics. To deliver the localisation or composition property, the sentence semantics should be able to present separation or disentanglement under connector $\oplus$. E.g., replacing *ARG0-animals* with *ARG0-fishes*.

**Formal semantic features in vector space.** After defining the semantic features of sentences, we propose the concept of a *convex cone of semantic feature*. In linear algebra, a *cone* refers to a subset of a vector space that is convex if any $\alpha \vec{v_i} + \beta \vec{v_j}$ if any $\vec{v_i}$ and $\vec{v_j}$ belong to it. $\alpha$ and $\beta$ are positive scalars. Formally, the definition of convex cone, $C$, is described as a set of vectors: $C = \{x \in V | x = \sum_{i=1}^{n} \alpha_i v_i, \alpha_i \geq 0, v_i \in R\}$ where $x$ is an element vector in vector space $\mathbb{R}$, $v_i$ are the basis vectors. $\alpha_i$ are non-negative scalars. In this context, we consider each *role-content* feature as a convex cone, $C$, corresponding to a hypersolid in high-dimensional vector space: $C_{c_i,r_i} = \{t(c_i, r_i) | t(c_i, r_i) \in sem(s), s \in corpus\}$ where $t(c_i, r_i)$ represents the basis vector in $C_{c_i,r_i}$ (Figure 2). According to set theory, we can define the formal semantic space as follows:

***Assumption1:*** *The sentence semantic space is the union of all unique $C_{c_i,r_i}$ convex cones:*

$$C_{c_1,r_1} \cup C_{c_2,r_2} \cup \cdots \cup C_{c_{V(c)},r_{V(r)}}$$

$V$ is the vocabulary of a corpus. Based on Assumption1, we can establish:

***Proposition1:*** *The geometrical location of sentence semantic vectors, $sem(s)$, can be determined by the intersection of different $C_{c_i,r_i}$:*

$$\begin{aligned} sem(s) &= t_1(c_1, r_1) \oplus \cdots \oplus t_i(c_i, r_i) \\ &= \{t_1(c_1, r_1)\} \oplus \cdots \oplus \{t_i(c_i, r_i)\} \\ &= C_{c_1,r_1} \cap C_{c_2,r_2} \cap \cdots \cap C_{c_i,r_i} \end{aligned}$$

## 4 Geometrical Formal Semantic Control

In this section, we first show that our formal semantic geometry can interpret sentence generation, such as arithmetic (Shen et al., 2020), and extend the "Linear Representation Hypothesis". Then, we propose a new semantic control approach, which recursively traverses the latent dimensions to probe the semantic geometry over latent spaces.

**Geometrical algebra interpretability.** Arithmetic has been considered a common way to control word or sentence semantics over latent spaces (Mikolov et al., 2013b). E.g., the addition operation can steer the sentence semantics (Shen et al., 2020; Mercatali and Freitas, 2021; Liu et al., 2023b), or linear interpolation can generate smooth intermediate sentences (Hu et al., 2022). However, they lack an explanation for these phenomena. We show that our geometrical framework can provide an intuitive explanation for these phenomena.

For linear interpolation, for example, it takes two sentences $x_1$ and $x_2$ and obtains latent vectors $z_1$ and $z_2$, respectively. It interpolates a path $z_k = z_1 \cdot (1-k) + z_2 \cdot k$ with $k$ increased from 0 to 1 by a step size of 0.1. Given two sentences with one role-content set overlap, $C_{c_j,r_j}$. We can describe:

$$
\begin{aligned}
&sem(s_1) \cap sem(s_2) \\
&= \{C^{s_1}_{c_1,r_1} \cap \dots \cap C^{s_1}_{c_i,r_i}\} \cap \{C^{s_2}_{c_1,r_1} \cap \dots \cap C^{s_2}_{c_i,r_i}\} \\
&= \{C^{s_1}_{c_1,r_1} \cap \dots \cap C^{s_2}_{c_i,r_i}\} \cap C^{s_{1(2)}}_{c_j,r_j}
\end{aligned}
$$

According to the definition of convex cone, if $z_1$ and $z_2$ are left in $C^{s_{1(2)}}_{c_j,r_j}$, the weighted sum vector, $z_t$, is also in $C^{s_{1(2)}}_{c_j,r_j}$. Therefore, the intermediate sentence semantics can be described as:

$$
\begin{aligned}
&sem(s^t_{1\to 2}) \\
&= (1-k) \times sem(s_1) + k \times sem(s_2) \\
&= \{\{z_1 \cdot (1-k) + z_2 \cdot k\}, \dots \{\dots\}\} \cap C^{s_{1(2)}}_{c_j,r_j}
\end{aligned}
$$

That is, the intermediate sentences will hold the $\{c_j, r_j\}$ information during interpolation.

**Linear representation hypothesis.** "Linear representation hypothesis" refers to high-level concepts being represented linearly as directions in representation space, which has been widely evaluated to interpret Large LMs' mechanism (Marks and Tegmark, 2023; Xie et al., 2021; Wang et al., 2024; Jiang et al., 2024b; Park et al., 2023, 2024). However, a main challenge for this hypothesis is that it's not clear what constitutes a high-level concept.

Our geometrical framework can further support and extend this hypothesis by answering the questions: What and how are they "linearly" encoded?

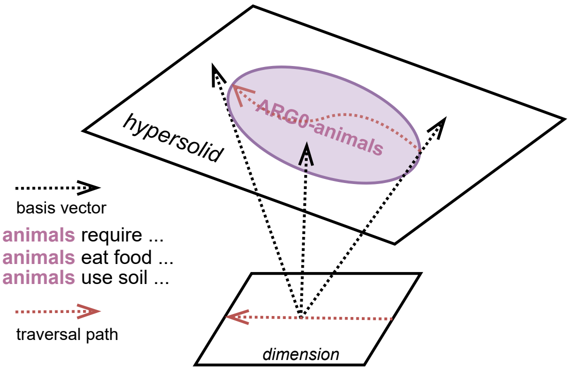



Figure 2: Algorithm 1: by modifying the latent dimensions, we can control the movement of latent vectors over latent space.

For example, given a set of $N$ atomic sentences: $s_i$: *bird is a kind of living thing* varying the content of arg1. Their semantics can be described below:

$$
sem(s) = \{C^{s_i}_{c_i,arg1}, \dots\} \cap \dots \cap C_{living\ thing,arg2}
$$

, where $c_i \in \{\text{tiger}, \text{bird}, \dots\}$

In this case, the concept *living thing* is encoded as a convex cone where all different $C^{s_i}_{c_i,arg1}$ contribute to its boundary, leading to a direction. The hierarchical relations between *living thing* and *bird, etc.* are determined by the convex cones *is a kind of*.

**Guided traversal.** Since we describe different sentence semantic features, $\{c_i, r_i\}$, as distinct convex cones, $C_{c_i,r_i}$, within a $N$-dimensional vector space, $V \in \mathbb{R}^N$, we can linearly divide each basis dimension, $i \in \{1, \dots, N\}$, into different value regions, $[a, b]^{(i)}$, based on minimal information entropy. Consequently, there is a sequence of dimensional subspaces for each semantic feature. Thus, movement between different $C_{c_i,r_i}$ regions can be achieved by moving out the dimensional regions within this sequence. This process can be implemented via a decision tree. In figure 3, for example, we can move the sentence from $C_{pred,causes}$ to $C_{pred,means}$ by modifying the values started from *dim 21* $\leq -0.035$, ..., ending at *dim 10* $\leq -1.11$. By traversing the tree path, we can control the sentence generation by moving between convex cones, detailed in Algorithm 1.

Based on our algorithm, we can use classification metrics as proxy metrics to evaluate latent space geometry. E.g., accuracy and recall for measuring feature *separability* and *density*.

**Algorithm 1** Guided latent space traversal

Datasets: $D = \{s_1, \ldots, s_n\}$
Labels: $Y = \{y_1, \ldots, y_n\}$, $y_i \in \{0, 1\}$
*# 0:pred-causes, 1:pred-means*
Seed: $s$ = *fire causes chemical change*
**for** $s_i \in D$ **do**
  $z_i \leftarrow \text{Encoder}(s_i)$
**end for**
$X \leftarrow \{z_1, \ldots, z_n\}$
tree $\leftarrow$ DecisionTreeClassifier$(X, Y)$
path $\leftarrow$ filter(tree) *# choose the shortest path between $C_0$ and $C_1$*
$z \leftarrow \text{Encoder}(s)$
**for** node $\in$ path **do**
  (dim, range, yes/no) $\leftarrow$ node
  **if** in current branch **do**
    z[dim] $\leftarrow v \notin$ range **if** yes **else** $v \in$ range
  **else do**
    z[dim] $\leftarrow v \in$ range **if** yes **else** $v \notin$ range
**end for**
$s \leftarrow$ Decoder(z) *# fire means chemical change*

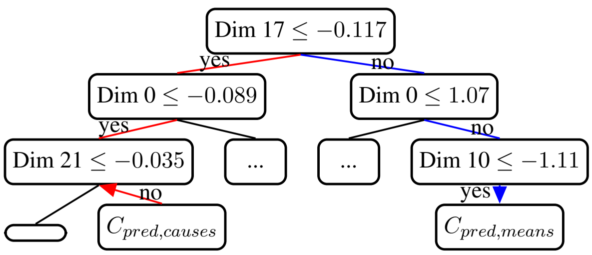



Figure 3: Traversal between different role-content sets by moving along the tree path.

## 5 SRL-Conditional VAE

In this section, we investigate the architecture of VAE to integrate the latent sentence space with LMs and propose a supervision approach to learn formal semantic geometry (i.e., role-content).

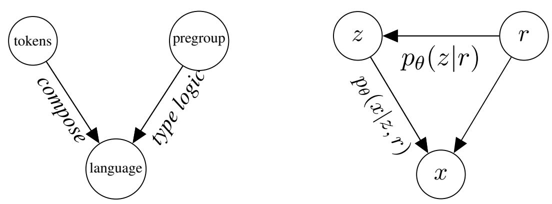
**Model architecture.** We consider Optimus (Li et al., 2020b) as the foundation which used BERT and GPT2 as Encoder and Decoder, respectively. In detail, the sentence representation, Embed(x), encoded from BERT[cls] will first transform into a Gaussian space by learning the parameters $\mu$ and $\sigma$ through multilayer perceptions $W_\mu$, $W_\sigma$. The final latent sentence representations can be obtained via: $z = W_\mu \times \text{Embed(x)} + W_\sigma$, which, as an additional Key and Value, is concatenated into the original Key and Value weights of GPT2, which can be described as: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{Q[z;K]^T}{\sqrt{d}})[z;V]$ where $Q$ has the shape $\mathbb{R}^{\text{seq}\times 64}$, $K, V$ has the shape $\mathbb{R}^{(\text{seq}+1)\times 64}$ (64 is dimension of GPT2 attention, seq is sequence length). Since $Q$ represents the target, $K$ and $V$ represent the latent representations. By intervening the $KV$ with $z$, we can learn the transformation between latent space and observation distribution.

**Optimisation.** It can be trained via the evidence lower bound (ELBO) on the log-likelihood of the data $x$ (Kingma and Welling, 2014). To bind the word content and semantic role information in latent space, we conditionally inject the semantic role sequence into latent spaces where the latent space $z$ and semantic role $r$ are dependent. The joint distribution can be described as:

$$P_\theta(x, r, z) = \underbrace{P_\theta(x|z, r)}_{\textit{likelihood}} \times \underbrace{P_\theta(z|r)}_{\textit{prior}} \times P(r)$$

Specifically, we first model the categorical structures by encoding the semantic roles sequence to learn the prior distribution with parameters $\mu^{(srl)}$ and $\sigma^{(srl)}$. Then, we jointly encode semantic roles and lexical tokens to learn the approximate posterior parameterised by $\mu$ and $\sigma$. By minimising the Kullback-Leibler (KL) divergence between prior and approximate posterior, the semantic features can be encoded in the latent sentence space. Moreover, to avoid the KL vanishing problem, which refers to the KL term in the ELBO becomes very small or approaching zero, we select the cyclical schedule to increase weights of KL $\beta$ from 0 to 1 (Fu et al., 2019) and a KL thresholding scheme (Li et al., 2019) that chooses the maximum between KL and threshold $\lambda$. The final objective function can be described as follows:

$$\mathcal{L}_{\text{CVAE}} = -\mathbb{E}_{q_\phi(z|r,x)}\Big[\log p_\theta(x|z, r)\Big] + \beta \sum_i \max\left[\lambda, \text{KL}q_\phi(z_i|x, r)||p(z_i|r)\right]$$




(a) CDM (Clark et al., 2008) (b) SRL-Conditional VAE

Figure 4: Comparison between Compositional Distributional Model (CDM) (left) and SRL-Conditional VAE (right).

where $q_\phi$ represents the approximated posterior (i.e., encoder). $i$ is the $i$-th latent dimension.

## 6 Empirical analysis

In the experiment, we quantitatively and qualitatively evaluate the latent space geometry via geometrical probing approaches: (1) traversal, (2) arithmetic, and (3) guided traversal. All experimental details are provided in Appendix A.

### 6.1 Latent Traversal

**Qualitative evaluation.** Traversal refers to the random walk over latent space. It can be done by decoding the latent vector in which each dimension is resampled and other dimensions are fixed (Higgins et al., 2017; Kim and Mnih, 2018; Carvalho et al., 2023). Given a latent vector from a "seed" sentence, we can traverse its neighbours to evaluate the geometry. As illustrated in Table 1, those traversed sentences can hold the same content under different semantic roles as the input, such as *automobile* in *ARG1*, indicating *role-content* feature separation in latent spaces.

an automobile is a kind of vehicle

an automobile is a kind of moving object
an automobile is a kind of object

an airplane is a kind of vehicle
a car is a kind of vehicle

Table 1: Traversal showing held semantic factors in explanations corpus.

**Quantitative evaluation.** Next, we employ t-SNE (Van der Maaten and Hinton, 2008) to examine *role-content* features cluster and separation over latent space (i.e., *natural clustering property* (Bengio, 2013)). In the corpus, however, due to the small number of data points within each role-content cluster, t-SNE cannot capture the differences between clusters well, resulting in the visualized latent space not displaying good role-content separability (top in figure 5). Therefore, we increase the number of data points in different role-content clusters by traversing each and keeping those resulting data points with the same role-content. Then, we visualise the role-content cluster at the bottom of figure 5. We can find that the features are clustered and separated over the latent space. If this was not the case, after traversing the resulting vectors from the same role-content cluster, the visualization should show the same entanglement as the original datapoints distribution.

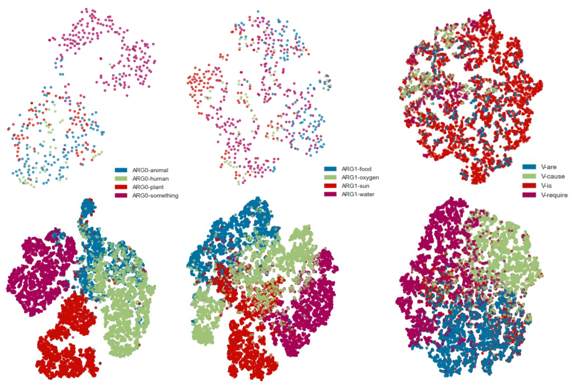


Figure 5: t-SNE plot of role-content distribution before and after traversal. From left to right are ARG0-(animal, human, plant, and something), ARG1-(food, oxygen, sun, and water), and predicate-(are, cause, is, require) (top: original role-cluster distribution, bottom: distribution after traversal). PCA plots are in Figure 9.

### 6.2 Latent Arithmetic

**Qualitative evaluation.** In addition, we demonstrate the geometric properties via interpolation in Table 2. For the top-most one, we can observe that sentences are smoothly moved from source to target (e.g., from *beach ball* to *atom* connected

a beach ball is a kind of container
1. a pool table is a kind of object
2. a balloon is a kind of object
3. a magnet is a kind of object
4. a neutron is a kind of particle
5. a proton is a kind of particle
an atom is a kind of particle

---

protons are found in the nucleus of an atom
1. protons are found in the nucleus of an atom
2. 1 atom is positive 1 in electric charge
3. 1 in 6000 is equal to 27 in 10 years
4. if protons and neutrons have the same number of neutrons then those two particles are physically closer than one another
5. if a neutron has a negative -10 electric charge then the atom will not be able to move
6. if a neutron has a negative -10 electric charge then the neutron will not have a positive electric charge
if a neutral atom loses an electron then an atom with a positive charge will be formed

Table 2: Interpolation examples (top: interpolation between sentences with similar semantic information, bottom: interpolation between sentences with different semantic information). Only unique sentences shown.

by *ballon*, *magnet*, *neutron*, and *proton*) where the same role-content (i.e., *pred-is*) unchanged. In contrast, the second case doesn't display the smooth interpolation path. E.g., the third sentence connecting different semantic structures is unrelated to both source and target due to a discontinuous space gap between different clusters. Both indicate that the explanatory sentences might be clustered according to different semantic role structures.

$s_1$: animals require food for survival
$s_2$: animals require warmth for survival
animals eat plants
animals produce milk
animals usually eat plants
animals eat berries ; plants
animals require food to survive
animals require shelter to survive

---

$s_1$: water vapor is invisible
$s_2$: the water is warm
igneous rocks are found under the soil
quartz is usually very small in size
quartz is formed by magma cooling
quartz is made of iron and zinc
silica is made of argon and argon
sedimentary is formed by lithosphere collapsing

Table 3: $s_1 \pm s_2$ (top: addition, bottom: subtraction).

Following the definition of convex cone, we next traverse the resulting sentence after adding or subtracting two sentences with the same role-content feature. As illustrated in Table 3, the adding operation tends to hold the same role-content (e.g., *ARG0-Animals*) as inputs. In contrast, the subtraction loses such control, e.g., from *ARG1-water* to *ARG1-quartz*. More similar observations are in Table 11. These results corroborate our geometry.

**Quantitative evaluation.** Next, we quantitatively assess our geometry framework by calculating the ratio of the same role-content results from the vector addition and subtraction for all sentence pairs with a matching role. As illustrated in Figure 6, the ADDed results (dark blue) can greatly hold the same token-level semantics (role-content) as inputs, indicating our geometrical framework. In contrast, the SUBed results (shallow blue) suffer from semantic shift. Similar observations for VERB and ARG1 can be found in Figure 11 and 12. Besides, we can quantify each role-content cluster's geometrical area by calculating the cosine similarity between randomly selected sentence pairs in this cluster. We report the maximal and

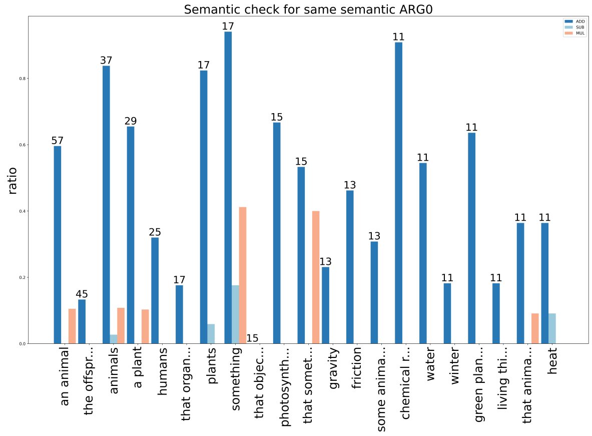



Figure 6: Arithmetic, $s_1 \pm s_2$, for ARG0 with contents (dark blue: addition, shallow blue: subtraction, orange: element-wise production).

minimal distance in Figure 7. Similar observations for VERB and ARG1 can be found in Figure 13 and 14.

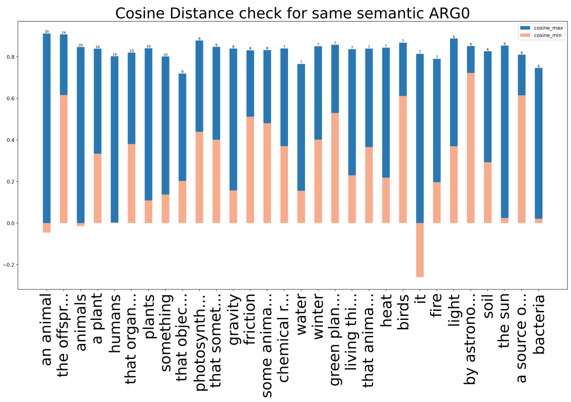



Figure 7: Evaluating the geometrical size of role-content clusters (blue: max, orange: min).

### 6.3 Guided Latent Traversal

Finally, we examine the semantic geometry via algorithm1. The categories selected below are chosen based on their frequencies, ensuring a balanced distribution during the classifier's training process.

**Qualitative evaluation.** Firstly, we evaluate the traversal between different semantic role structures, e.g, conditional and atomic sentences. Table 4 shows that the cluster of the generated sentence changes as the values of different dimensions change sequentially (e.g., the first three sentences hold the same characteristic *if ... then ...* as the input. The remaining sentences gradually move closer to the target characteristics, such as *is*). Meanwhile, the sentences can hold the subject, *something*, during the movement, corroborating our geometry framework. Next, we evaluate the

if something receives sunlight it will absorb the sunlight
Dim27: if a thing absorbs sunlight then that thing is warmer
Dim12: if something is eaten then that something produces heat
Dim08: if something gets too hot in sunlight then that something is less able to survive
Dim03: something contains physical and chemical energy
Dim21: something contains sunlight
Dim10: some things are made of matter
Dim00: something is made of atoms
Dim17: a forest contains life
Dim00: something that is cold has a lower temperature
Dim21: something rises in temperature
Dim00: something is formed from things dissolved in water
Dim30: something that is cold has fewer nutrients
Dim21: something that is not moved is dead

Table 4: Movement from *conditional* to *atomic* sentences.

traversal between predicates. Table 5 shows the movement between verbs (*cause* and *mean*). We can observe that the predicate is modified from *causes* to *mean*. In the traversal process, some sentences fall into the *V-is* region. The reason is that the *V-is* cluster is widely scattered in latent space (shown in Figure 5), which leads to a big overlap between *V-is* and *V-mean*. Moreover, we calculate the ratio of the generated sentences that hold the expected predicate, *mean*, from 100 sentences with predicate *cause*. The ratio is 0.71, which indicates that the decision tree is a reliable way to navigate the movement of sentences. Finally, we evaluate the traversal between arguments. Table 6 shows the movement from argument *water* to *something*. Similarly, the ARG1 can be modified from *water* to *something* following its path. Besides, the final generated explanation still holds a similar semantic structure, *is a kind of*, compared with input.

**Quantitative evaluation.** Finally, we use classification metrics, including accuracy (*separability*) and recall (*density*), as proxy metrics to assess latent space geometry. As shown in Table 7, all features show higher separation where argument1 leads to the highest separation, indicating latent space geometry.

fire causes chemical change
Dim06: fire causes chemical changes
Dim22: fire causes chemical reactions
Dim02: fire can cause harm to plants
Dim27: smoke can cause harm to organisms
Dim14: fire causes physical harm to objects
Dim24: fire can cause chemical changes
Dim08: fire destroys material
Dim01: fire means chemical change
Dim14: waste means igneous metal
Dim06: combustion means burning
Dim00: combustion means chemical changes
Dim21: combustion means burning
Dim00: fire is formed by thermal expansion
Dim18: fire chemical means chemical energy
Dim03: fire is corrosive

winter means cold environmental temperature
Dim03: winter means cold - weather
Dim18: winter means cold weather
Dim00: winter means weathering
Dim21: drought means high temperatures / low precipitation
Dim00: winter means high amounts of precipitation
Dim06: drought causes natural disasters
Dim14: drought has a negative impact on crops
Dim01: drought has a negative impact on animals
Dim08: drought causes animal populations to decrease
Dim24: drought causes ecosystem loss
Dim14: drought causes animals to have lower natural temperature
Dim27: cold climates causes wildfires
Dim02: climate change can cause low rainfall
Dim22: global warming causes droughts
Dim06: winter causes weather patterns

Table 5: Movement between *cause* and *mean*.

water is a kind of substance
Dim12: water is a kind of substance
Dim00: water is a kind of liquid
Dim23: liquid is a kind of material
Dim29: water has a positive impact on a process
Dim17: absorbing water is similar to settling
Dim06: absorbing is similar to reducing
Dim21: absorbing something is similar to absorbing something
Dim04: storing something means being protected
Dim06: producing something is a kind of process
Dim04: storing something is similar to recycling
Dim21: absorbing something is a kind of process
Dim01: absorbing something can mean having that something
Dim22: folding something is similar to combining something
Dim07: improving something is a kind of transformation
Dim11: absorbing something is a kind of method
Dim07: absorbing something is a kind of process

Table 6: Movement from *water* to *something*.

| Formal semantic features | separation↑ | density↑ |
|---|---|---|
| predicate (causes, means) | 0.87 | 0.92 |
| argument1 (water, something) | 0.95 | 0.48 |
| structure (condition, atomic) | 0.58 | 0.55 |

Table 7: Proxy metrics for latent space geometry.

## 7 Conclusion and Future Work

In this study, we investigate the localisation of general semantic features to enhance the controllability and explainability of distributional space from the perspective of formal semantics, which is currently under-explored in the NLP domain. We first propose the formal semantic features as *role-content* and define the corresponding geometrical framework. Then, we propose a supervision approach to bind the semantic role and word content. In addition, we propose a novel traversal probing approach to assess the latent space geometry based on information set and entropy. We extensively evaluate the latent space geometry through geometrical operations, such as traversal, arithmetic, and our guided traversal. Experimental results indicate the existence of formal semantic geometry.

Since recent theoretical works reveal that the LLMs can encode linear symbolic concepts (Jiang et al., 2024b), in the future, we will explore their in-context learning of compositional semantics based on our formal semantic geometry framework.

## 8 Limitations

**1.** Limitation of data source: this work only focused on explanatory sentences. Whether the semantic separability of other corpora emerges over latent space is not explored. **2.** Role-content clusters overlapping: the geometric analysis indicates that the role-content regions still have significant overlapping over distributional spaces. Therefore, a new potential task can be how we can better separate/disentangle the semantic features (role-content) to provide better localisation or composition behaviour over distributional semantic spaces in the Computational Linguistics domain, further assisting downstream tasks, such as Natural Language Reasoning, Compositional Generalisation, etc. **3.** Large Language Models: this paper only investigates the BERT-GPT2 architecture based on the current state-of-the-art language VAE (Optimus). The larger decoder is out of the scope of this work and needs to be investigated in the future.


## Acknowledgements

We appreciate the reviewers for their insightful comments and suggestions. This work was partially funded by the EPSRC grant EP/T026995/1 entitled “EnnCore: End-to-End Conceptual Guarding of Neural Architectures” under Security for all in an AI enabled society, by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617), by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre and the NIHR Manchester Biomedical Research Centre.


## References


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

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019.

Yoshua Bengio. 2013. Deep learning of representations: Looking forward. In *International conference on statistical language and speech processing*, pages 1–37. Springer.

Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, and Greg Durrett. 2021. Flexible generation of natural language deductions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6266–6278, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danilo S. Carvalho, Yingji Zhang, Giangiacomo Mercatali, and Andre Freitas. 2023. Learning disentangled representations for natural language definitions. *Findings of the European chapter of Association for Computational Linguistics (Findings of EACL)*.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. 2022. The geometry of multilingual language model representations. *arXiv preprint arXiv:2205.10964*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. 2008. A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140. Oxford.

Stephen Clark and Stephen G. Pulman. 2007. Combining symbolic and distributional models of meaning. In *Quantum Interaction*.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees.

Donald Davidson. 1967. The logical form of action sentences.

David R Dowty, Robert Wall, and Stanley Peters. 2012. *Introduction to Montague semantics*, volume 11. Springer Science & Business Media.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2020. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Ghazi Felhi, Joseph Le Roux, and Djamé Seddah. 2022. Towards unsupervised content disentanglement in sentence representations via syntactic roles. *arXiv preprint arXiv:2206.11184*.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. 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)*, pages 240–250, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiyan Fu and Anette Frank. 2024. Exploring continual learning of compositional generalization in nli. *arXiv preprint arXiv:2403.04400*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson H S Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2017. A deep semantic natural language processing platform.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, page 512–520, USA. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. Controllable text generation via probability density estimation in the latent space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12590–12616, Toronto, Canada. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. 2022. Fuse it more deeply! a variational transformer with layer-wise latent variable inference for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 697–716, Seattle, United States. Association for Computational Linguistics.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Ray S Jackendoff. 1992. *Semantic structures*, volume 18. MIT press.

Peter Jansen, Rebecca Sharp, Mihai Surdeanu, and Peter Clark. 2017. Framing qa as building and ranking intersentence answer justifications. *Computational Linguistics*, 43(2):407–449.

Peter A Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *arXiv preprint arXiv:1802.03052*.

Giyoung Jeon, Haedong Jeong, and Jaesik Choi. 2019. An efficient explorative sampling considering the generative boundaries of deep generative neural networks.

Yibo Jiang, Bryon Aragam, and Victor Veitch. 2024a. Uncovering meanings of embeddings via partial orthogonality. *Advances in Neural Information Processing Systems*, 36.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. 2024b. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. *arXiv preprint arXiv:1704.05572*.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes.

Robert B Lees. 1957. Syntactic structures.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. 2019. A surprisingly effective fix for deep latent variable modeling of text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3603–3614, Hong Kong, China. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020b. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022a. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Zhuang Li, Lizhen Qu, Qiongkai Xu, Tongtong Wu, Tianyang Zhan, and Gholamreza Haffari. 2022b. Variational autoencoder with disentanglement priors for low-resource task-specific natural language generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10335–10356, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William W Cohen. 2020. Differentiable open-ended commonsense reasoning. *arXiv preprint arXiv:2010.14439*.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2023a. Composable text controls in latent space with ODEs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16543–16570, Singapore. Association for Computational Linguistics.

Sheng Liu, Lei Xing, and James Zou. 2023b. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. 2021. Smoothing the disentangled latent style space for unsupervised image-to-image translation.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Giangiacomo Mercatali and André Freitas. 2021. Disentangling generative factors in natural language with discrete variational autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3547–3556.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023. Language models implement simple word2vec-style vector arithmetic. *arXiv preprint arXiv:2305.16130*.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.

Melanie Mitchell. 2023. How do we know how smart ai systems are?

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*.

Martha Stone Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic role labeling*, volume 6. Morgan & Claypool Publishers.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2024. The geometry of categorical and hierarchical concepts in large language models.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Malka Rappaport Hovav and Beth Levin. 2008. The english dative alternation: The case for verb sensitivityl. *Journal of linguistics*, 44(1):129–167.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising. In *International conference on machine learning*, pages 8719–8729. PMLR.

Paul Smolensky. 2006. Harmony in linguistic cognition. *Cognitive science*, 30(5):779–801.

Paul Smolensky and Géraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar. Vol. 1, Cognitive architecture*. MIT.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Explainable inference over grounding-abstract chains for science questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1–12.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2024. A differentiable integer linear programming solver for explanation-based natural language inference. *arXiv preprint arXiv:2404.02625*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. 2023. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Asahi Ushio, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? *arXiv preprint arXiv:2105.04949*.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022a. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11403–11411.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2020. Explainable natural language reasoning via conceptual unification. *arXiv preprint arXiv:2009.14539*.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2022b. Case-based abductive natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1556–1568.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Emile van Krieken, Thiviyan Thanapalasingam, Jakub Tomczak, Frank Van Harmelen, and Annette Ten Teije. 2023. A-nesi: A scalable approximate method for probabilistic neurosymbolic inference. *Advances in Neural Information Processing Systems*, 36:24586–24609.

Jake Vasilakes, Chrysoula Zerva, Makoto Miwa, and Sophia Ananiadou. 2022. Learning disentangled representations of negation and uncertainty. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8380–8397, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2024. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. SyGNS: A systematic generalization testbed based on natural language semantics. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 103–119, Online. Association for Computational Linguistics.

## A Experiment Setting

**Dataset.** Table 8 displays the statistical information of the datasets used in the experiment. The data of the two datasets partially overlap, so only the unique explanations are selected as the experimental data. The rationale for choosing explanatory sentences is that they are designed for formal/localised/symbolic semantic inference task in natural language form, which provides a semantically complex and yet controlled experimental setting, containing a both well-scoped and diverse set of target "concepts" and sentence structures, providing a semantically challenging yet sufficiently well-scoped scenario to evaluate the syntactic and semantic organisation of the space. Besides, those concepts mentioned in the corpus, such as *animal is a kind of living thing*, are fundamental to human semantic understanding.

| Corpus | Num data. | Avg. length |
|---|---|---|
| WorldTree (Jansen et al., 2018) | 11430 | 8.65 |
| EntailmentBank (Dalvi et al., 2021) | 5134 | 10.35 |

Table 8: Statistics from explanations datasets.

Table 9 illustrates the semantic, structure, and topic information of explanatory sentences over the latent space. The explanatory sentences are automatically annotated using the semantic role labelling (SRL) tool, which can be implemented via AllenNLP library (Gardner et al., 2017). We report in Table 10 the semantic roles from the explanations corpus.

**Architecture.** Figure 8 provides a visual representation of the connection between BERT and GPT2 within the AutoEncoder architecture.

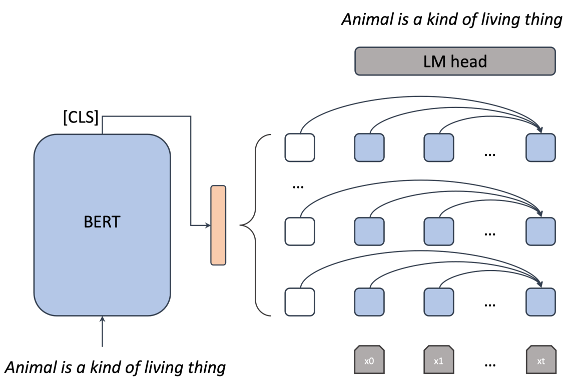



Figure 8: Latent sentence injection.

To train the CVAE, we use a new embedding layer for semantic roles and separate MLP layers $W_{\mu}^{srl}$ and $W_{\sigma}^{srl}$ to learn prior distribution.

**Hyperparameters.** The training process of the decision tree binary classifier can be implemented via scikit-learn packages with default hyperparameters. As for Optimus, the latent space size is 32 in the experiment. The training details are following the original experiment from Optimus (Li et al., 2020b).

## B Further Experimental Results

**Traversal visualisation.** PCA plots for ARG0, ARG1, and PRED are provided in Figure 9.

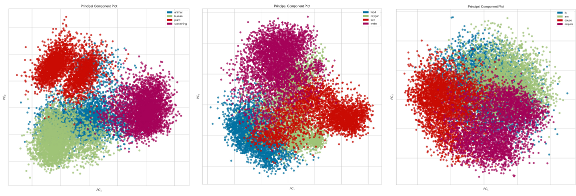


Figure 9: PCA visualisation.

In addition, we also provide the visualisation of word content *animal* with different semantic roles: ARG0, ARG1, ARG2, in Figure 10. From it, we can observe that the same content with different semantic roles can also be clustered and separated in latent space.

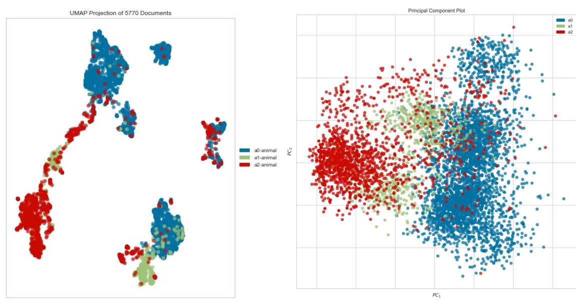


Figure 10: Visualisation for *animal-ARG0,1,2*.

**Qualitative evaluation for arithmetic.** Table 11 lists the traversed explanations after addition (blue) and subtraction (red) on different semantic role information. We can observe that the resulting sentences after addition can hold the same role-content as inputs, revealing latent space geometry.

**Quantitative evaluation for arithmetic.** Quantitative evaluation for our hypotheses via latent arithmetic. Both VERB and Object can perform high

| Cluster | Theme and Pattern |
|---|---|
| 0 | Theme: physics and chemistry. Pattern: *if then* and *as*. E.g., if a substance is mixed with another substance then those substances will undergo physical change. |
| 1 | Theme: country, astronomy, and weather. E.g., new york state is on earth |
| 2 | Theme: physics and chemistry. Pattern: *is a kind of*. E.g., light is a kind of wave. |
| 3 | Theme: biology. E.g., a mother births offspring. |
| 4 | Theme: synonym for verb. Pattern: *means* and *is similar to*. E.g., to report means to show. |
| 5 | Theme: astronomy. E.g., the solar system contains asteroids. |
| 6 | Theme: animal/plant. Pattern: *is a kind of*. E.g., a seed is a part of a plant. |
| 7 | Theme: item. E.g., a telephone is a kind of electrical device for communication. |
| 8 | Theme: synonym for life. Pattern: *means* and *is similar to*. E.g., shape is a kind of characteristic. |
| 9 | Theme: geography. Pattern: *is a kind of*. E.g., a mountain is a kind of environment. |
| 10 | Theme: animal and plant. Pattern: *if then* and *as*. E.g., if a habitat is removed then that habitat is destroyed. |
| 11 | Theme: scientific knowledge. Pattern: *(;)*, *number* and */*. E.g., freezing point is a property of a ( substance ; material ). |
| 12 | Theme: item. Pattern: *is a kind of object*. E.g., a paper is a kind of object. |
| 13 | Theme: chemistry and astronomy. E.g., oxygen gas is made of only oxygen element. |
| 14 | Theme: general about science. Pattern: *(;)*. E.g., seed dispersal has a positive impact on ( a plant ; a plant 's reproduction). |
| 15 | Theme: item. Pattern: *is a kind of*. E.g., fertilizer is a kind of substance. |
| 16 | Theme: physics and chemistry. Pattern: *(;)*. E.g., the melting point of oxygen is -3618f ; -2188c ; 544k. |
| 17 | Theme: animal. E.g., squirrels live in forests. |
| 18 | Theme: nature. E.g., warm ocean currents move to cooler ocean regions by convection. |
| 19 | Theme: life. E.g., pond water contains microscopic living organisms. |

Table 9: Cluster Information for explanatory sentences, we use a k-means classifier to classify the sentence representations and manually evaluate each class.

| Semantic Tags | Prop. % | Description and Example |
|---|---|---|
| ARGM-DIR | 0.80 | Directionals. E.g. all waves transmit energy **from one place to another** |
| ARGM-PNC | 0.08 | Purpose. E.g. many animals blend in with their environment **to not be seen by predators** |
| ARGM-CAU | 0.05 | Cause. E.g. cold environments sometimes are white in color **from being covered in snow** |
| ARGM-PRP | 1.30 | Purpose. E.g. a pot is made of metal **for cooking** |
| ARGM-EXT | 0.04 | Extent. E.g. as the amount of oxygen exposed to a fire increases the fire will burn **longer** |
| ARGM-LOC | 4.50 | Location. E.g. a solute can be dissolved **in a solvent** when they are combined |
| ARGM-MNR | 2.00 | Manner. E.g. fast means **quickly** |
| ARGM-MOD | 9.80 | Modal verbs. E.g. atom **can** not be divided into smaller substances |
| ARGM-DIS | 0.07 | Discourse. E.g. if something required by an organism is depleted **then** that organism must replenish that something |
| ARGM-GOL | 0.20 | Goal. E.g. We flew **to Chicago** |
| ARGM-NEG | 1.20 | Negation. E.g. cactus wrens building nests in cholla cacti does **not** harm the cholla cacti |
| ARGM-ADV | 6.70 | Adverbials |
| ARGM-PRD | 0.20 | Markers of secondary predication. E.g. |
| ARGM-TMP | 7.00 | Temporals. E.g. a predator **usually** kills its prey to eat it |
| O | - | Empty tag. |
| V | 100 | Verb. |
| ARG0 | 32.0 | Agent or Causer. E.g. **rabbits** eat plants |
| ARG1 | 98.5 | Patient or Theme. E.g. rabbits eat **plants** |
| ARG2 | 60.9 | indirect object / beneficiary / instrument / attribute / end state. E.g. animals are **organisms** |
| ARG3 | 0.60 | start point / beneficiary / instrument / attribute. E.g. sleeping bags are designed **to keep people warm** |
| ARG4 | 0.10 | end point. E.g. when water falls from the sky that water usually returns **to the soil** |

Table 10: Semantic Role Labels that appears in explanations corpus.

ratio after addition, indicating role-content separability.

**ADD and SUB arithmetic**

ARGUMENT1:
a needle is a kind of object
a tire is a kind of object

a wire is a kind of object
a stick is a kind of object
a ball is a kind of object

a serotype is similar to intersex egg
a zygote contains many cell types
an xylem is made of two clumps

VERB:
chromosomes are located in the cells
Australia is located in the southern hemisphere

stars are located in the solar system
Jupiter is located in the milky way galaxy
aurora is located in the constellation of Leo

a crystal is made of metal
an alloy is made of iron and zinc
an aluminum plug is nonmagnetic

LOCATION:
volcanoes are often found under oceans
mosquitos can sense carbon dioxide in the air

polar ice sheets are located along rivers
hurricanes occur frequently along the coast in Africa
tide waves cause flooding in coastal waters

valley is a kind of location
shape is a property of rocks
desert is a kind of place

TEMPORAL:
as the population of prey decreases competition between predators will increase
as competition for resources decreases the ability to compete for resources will increase

as the population of an environment decreases ecosystem function will decrease
as the spread of available air mass increases the population will increase
as the number of heavy traffic required increases the traffic cycle will decrease

some types of lizards live in water
a rose is rich in potassium
a fern grass roots foot trait means a fern grass

NEGATION:
pluto has not cleared its orbit
sound can not travel through a vacuum

radio waves don't have electric charge
electromagnetic radiation does not have a neutral electric charge
electromagnetic radiation contains no electric charge

Mars is a kind of moon / planet
Anothermic rock is a kind of metamorphic rock
Anal Cetus's skeleton is a kind of fossil

Table 11: Latent sapce arithmetic for five semantic tags (blue: addition, red: subtraction).

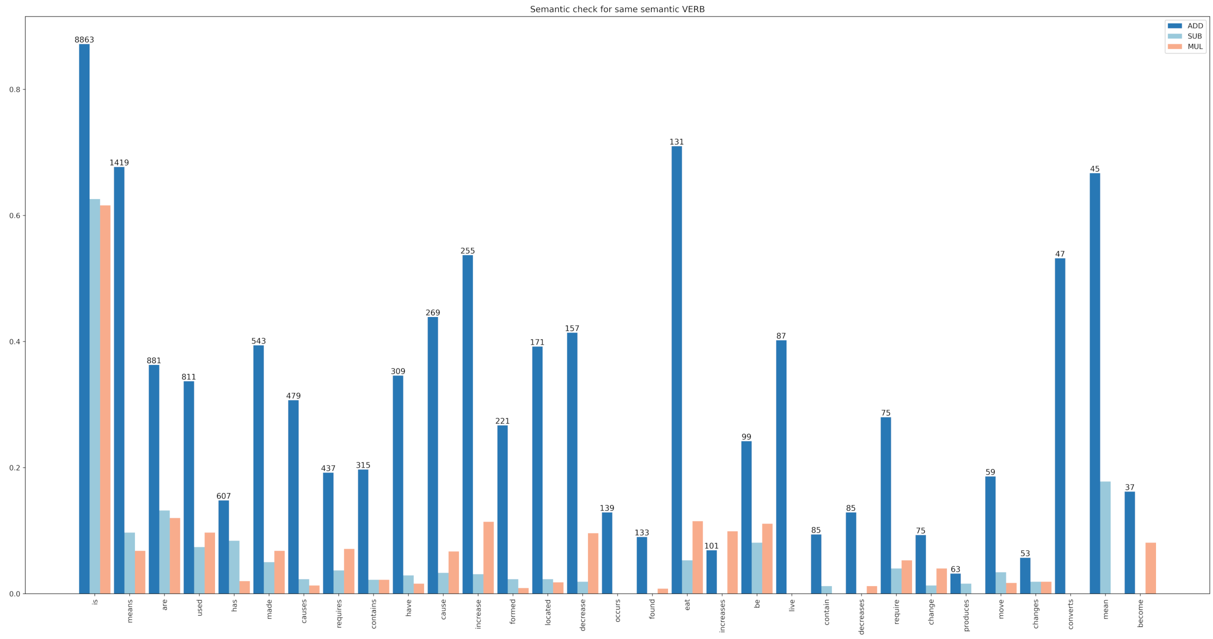



Figure 11: Predicate (VERB). The content *is* shows the high ratio after subtraction, indicating that the *V-is* is widely distributed over the latent space.

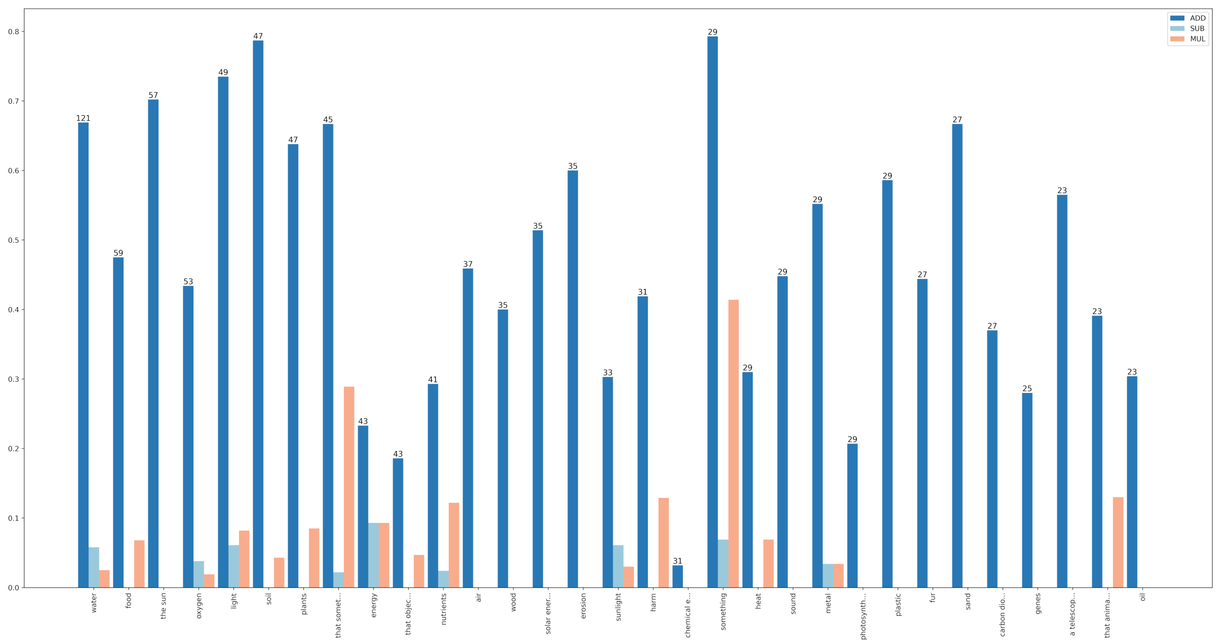



Figure 12: Object (ARG1).

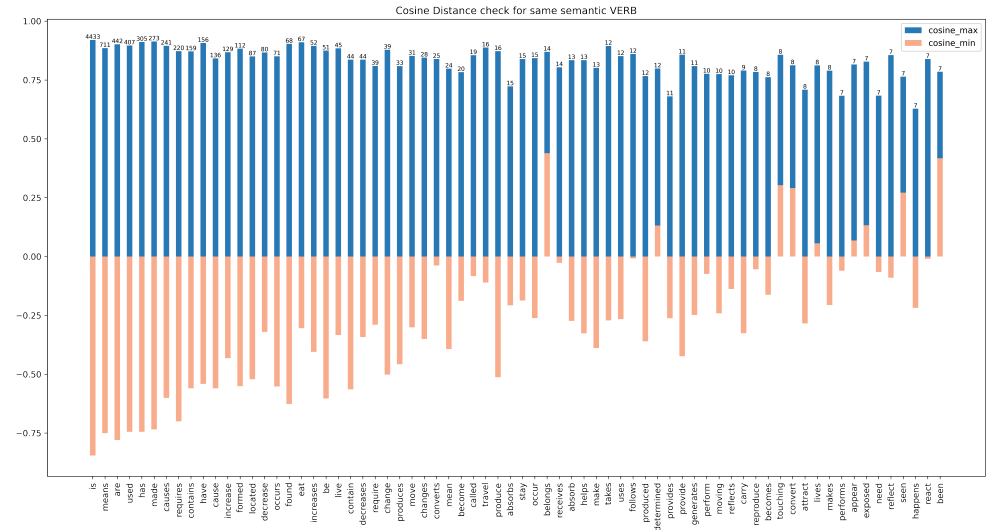



Figure 13: Cosine distance of sentence pairs in VERB-content clusters.

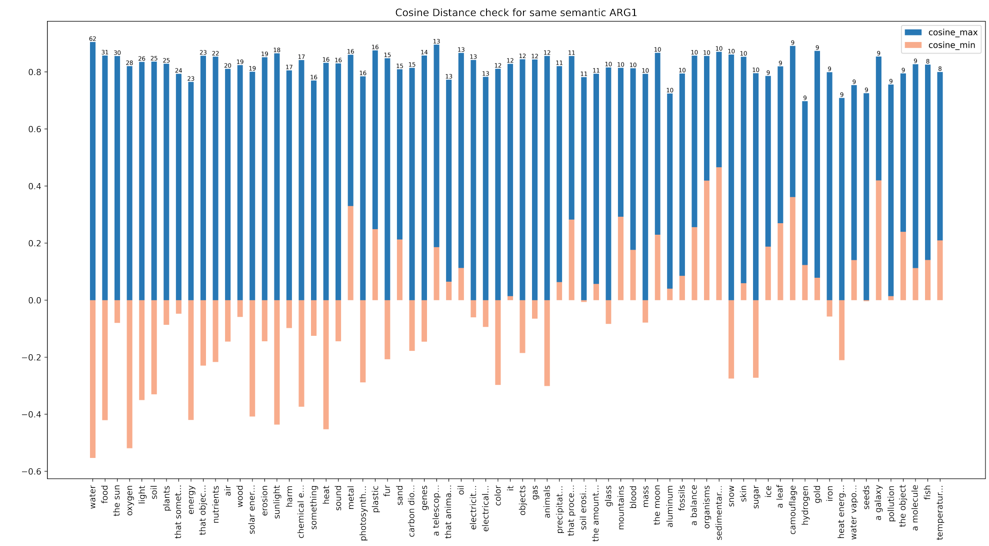



Figure 14: Cosine distance of sentence pairs in ARG1-content clusters.