Learning Disentangled Semantic Spaces of Explanations via Invertible Neural Networks

Anonymous ACL submission

Abstract

Disentangled latent spaces usually have better semantic separability and geometrical properties, which leads to better interpretability and more controllable data generation. While this has been well investigated in Computer Vision, in tasks such as image disentanglement, in the NLP domain sentence disentanglement is still comparatively under-investigated. Most previous work have concentrated on disentangling task-specific generative factors, such as sentiment, within the context of style transfer. In this work, we focus on a more general form of sentence disentanglement, targeting the localised modification and control of more general sentence semantic features. To achieve this, we contribute to a novel notion of sentence semantic disentanglement and introduce a flow-based invertible neural network (INN) mechanism integrated with a transformer-based language Autoencoder (AE) in order to deliver latent spaces with better separability properties. Experimental results demonstrate that the model can conform the distributed latent space into a better semantically disentangled sentence space, leading to improved language interpretability and controlled generation when compared to the recent state-of-the-art language VAE models.

1 Introduction

Most previous work on controlled text generation have concentrated on style transfer tasks: modifying sentences with regard to markers of sentiment, formality, affirmation/negation (John et al., 2019; Bao et al., 2019; Hu and Li, 2021; Vasilakes et al., 2022; Gu et al., 2022; Liu et al., 2023; Gu et al., 2023) (Figure 1 top). Disentanglement of language generative factors over Variational Autoencoder (VAE) spaces has been a key mechanism to deliver this type of generative control (John et al., 2019; Bao et al., 2019; Vasilakes et al., 2022). Recently, Zhang et al. (2022) demonstrated that a more general form of semantic control can be achieved in the latent space of Optimus (Li et al., 2020b), the first standard transformer-based VAE, where a BERT (Devlin et al., 2018) encoder and a GPT2 (Radford et al., 2019) decoder are connected within a VAE bottleneck. Using representations of conceptually dense *explanatory sentences* (Jansen et al., 2018b), they showed that sentences (e.g. *animals require oxygen for survival*), can be represented within a space which can be organised around the associations between predicate, arguments and their associated token content: *ARG0-animals* or *VERB-require*, is geometrically resolved to a hypersolid over the latent space. Nevertheless, the ability to

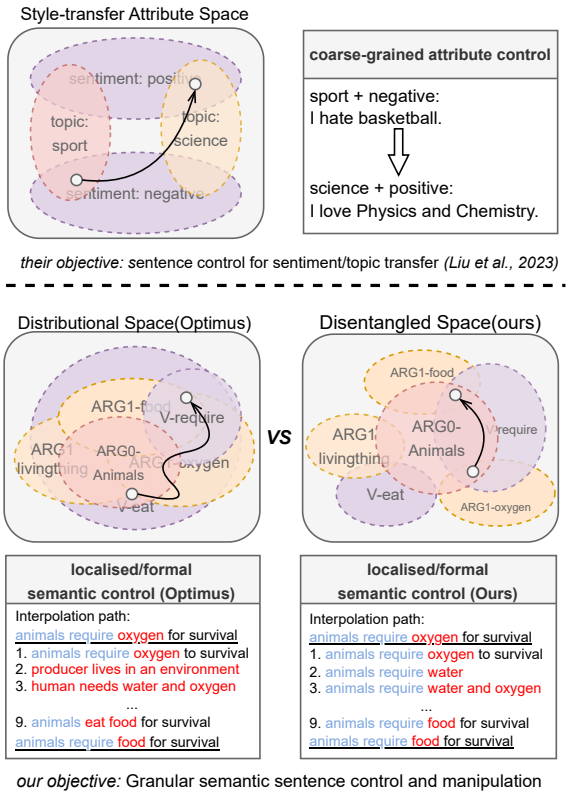

Figure 1: Top: attribute space geometry. Bottom: general semantic geometry, where left: distributional semantic space of Optimus (Li et al., 2020b), right: our compositionality-induced semantic space where the geometrical location of sentence vectors can be located by the intersection of role-content clusters.

learn and control such separation is still limited as different semantic factors of the sentence are still overlapped and entangled in the latent space (e.g., *V-eat* and *V-require* in Figure 1 bottom left), indicating distributional sentence semantics cannot be currently localised and controlled from the perspective of formal semantics (i.e., *predicate-argument structures*, *compositionality*) (Marcus, 2003; Nefdt, 2020; Dankers et al., 2022).

This work aims to improve the localisation and semantic control of latent sentence spaces, by delivering a model which can better separate and control syntactic-semantic features (e.g. predicate-argument) and their associated lexical semantics content. This type of representation can provide the foundation 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.

To deliver this type of semantic control within the distributional sentence space, following the methodological framework introduced by (Zhang et al., 2022), we target on improving the semantic separability of sentences by focusing on explanatory sentences [1], rather than synthetic or style transfer datasets (Hupkes et al., 2020; Yanaka et al., 2021), in which the *semantic structure of sentences* can be isolated and controlled. Inspired by the work of Esser et al. (2020), we integrate a flow-based invertible neural network (INN) (Dinh et al., 2014) as a plug-in control component to learn the bijective transformation between the distributional hidden space of the transformer-based language autoencoder (BERT-GPT2) and the smooth Gaussian space of the INN (Figure 2). Specifically, we first pre-train an autoencoder (AE) to learn sentence representations from the transformers' latent spaces. Then, we freeze the AE weights and train the INN to map the AE representations to a Gaussian space. Since INN models define a bijective transformation, we can control the autoencoder generation by manipulating the INN latent spaces, which is more efficient and significantly less resource intensive than re-training a language AE end-to-end.

More importantly, we propose a supervised training strategy within the INN setting to learn a latent space with improved semantic separability, namely: the *semantic role-content pairs* and their associated clusters can be better separated over the latent space modelled by the INN (Section 4.1). In this case, we can improve localised control over the decoding process due to the reduction of overlapping (ambiguous) regions. *A more separable and geometrically consistent sentence space* can be then operated over to improve the generative control with support of geometric operators, such as interpolation (Bowman et al., 2016) (Section 4.2). The contributions of this work are summarised below:

1. We approach sentence disentanglement and generation control from the point of view of *Argument Structure Theory (AST)*, bridging latent space features with a canonical, linguistics-informed, semantic representation of sentences. **2.** We find that integrating a flow-based INN mechanism into a transformer-based language-AE architecture is an effective mechanism for transforming the hidden space of the autoencoder into a smooth Gaussian latent space for representing sentences. **3.** We propose a supervised training strategy for INNs to learn a controllable semantic space with higher disentanglement and separability of semantic features, when compared to previous work. **4.** Using this mechanism, we systematically employ geometrical data augmentation strategies to assist on sentence representation disentanglement.

Interpreting and controlling sentence generation from the perspective of the geometric manipulation of the latent space is still largely unexplored within NLP. To the best of our knowledge, this is the first work which focuses on the introduction of invertible NN-based mechanisms to support latent spaces with better separated argument structure/semantic features, allowing for a more universal form of sentence generation control. All code and supporting datasets are available online[2].

2 Preliminaries

In this section, we first define the sentence representation model based on Argument Structure Theory (AST), linking with the associated disentanglement/generative factors and then proceed with the description of the proposed flow-based INN mechanism.

Sentence semantic disentanglement. AST (Jackendoff, 1992; Levin, 1993; Rappaport Hovav and Levin, 2008) provides a model for representing sen-

[1] The rationale for choosing explanatory sentences and their semantic details are provided in Appendix A.

[2] https://github.com/<anonymized>

tence 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). Therefore, this work uses the notion of sentence semantic disentanglement as the cluster separation of the content under the PArg/SRs structure (the corresponding *role* in role-content), aiming to induce a latent space which geometrically encodes the AST structure, better disentangling and separating role-content clusters (Figure 1).

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

Formally, a sentence s (e.g., see above) consists of a sequence PArgs/SRs and word content associations. Upon encoding in latent space, this can be described as:

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

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 et al., 2008). $\oplus$: connects the lexical semantics (word content + structural role) to form the sentence semantics. This work applies distinct symbols aiming to emphasise the disentanglement aspects associated with the AST structure. If the sentence representation can be semantically disentangled under $\oplus$, the $sem(s)$ can be decomposed into: $sem(s) = \{t_1(c_1, r_1)\} \oplus \cdots \oplus \{t_i(c_i, r_i)\}$ where each set represents a specific role-content cluster resolved to a hypersolid over the latent space, in this case, given a set of N sentences within the same predicate cluster $t(c, r)$ (i.e., *V-require*) but different $sem(s)$, those sentence vectors can represent $t(c, r)$ features independently of other features (i.e., *ARG0-animals*), forming the $t(c, r)$ cluster: $\{sem(s_1), ..., sem(s_N)\} = \{t(c, r)\}_{\times N} \oplus \{...\}$ Therefore, we can evaluate the disentanglement (i.e., *natural clustering property* (Bengio, 2013)) of sentence semantics by evaluating the density within $\{t(c, r)\}$ set(cluster) (classifier recall) and the separation between different $\{t(c, r)\}$ set(clusters) (classifier accuracy) with downstream classifiers based on the *manifold hypothesis for classification* (Rifai et al., 2011), rather than disentanglement metrics, which usually calculate the separation between latent dimensions, commonly used in the image domain (Higgins et al., 2017; Kim and Mnih, 2018; Chen et al., 2018; Ridgeway and Mozer, 2018). Next, we will introduce the INN-based mechanism which is used to learn this semantically disentangled space.

Invertible Neural Networks (INNs). Flow-based INNs (Dinh et al., 2014, 2016) are a class of neural networks that models the bijective mapping between the observation distribution $p(x)$ and the latent distribution $p(z)$. We use T to represent the forward mapping (from $p(x)$ to $p(z)$) and T' to represent the backward mapping (from $p(z)$ to $p(x)$), respectively. Unlike VAEs that approximate the prior distribution to multivariate Gaussian distributions, INNs exactly use multivariate Gaussian distributions. These are trained by the following objective function: $\mathcal{L} = -\mathbb{E}_{x \sim p(x)} \Big[T(x) \Big]^2 - \log |T'(x)|$ where $T(x)$ learns the transformation from x to $z \sim N(0, 1)$. $|T'(x)|$ is the determinant of the Jacobian for $T(x)$, which indicates the extent in which the transformation locally expands or contracts the space. The term $-\log |T'(x)|$ ensures the integration of the probability density function to be one. The forward and reversed mapping can be implemented via the *coupling* layer (Dinh et al., 2014; Kingma and Dhariwal, 2018).

The rationale for choosing flow-based INNs lies on the fact that they learn the bijective transforma-

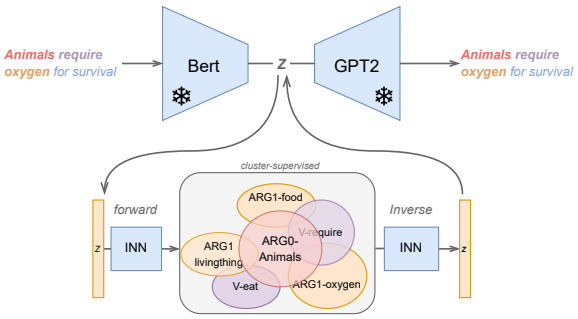

Figure 2: Transforming the representations of explanatory sentences from a language autoencoder (BERT-GPT2), into asemantically separable latent space with the support of the INN mechanism, where a sentence representation can be decomposed into a predicate-argument-level semantics (role-content).

tion between the latent and observed spaces, which can be used to guide the autoencoder generation by manipulating the INN latent space, which is more efficient and has lower computational demand than re-training a language VAE. Besides, flow-based INNs that learn the prior distribution (i.e., Gaussian) exactly, can theoretically prevent the information loss from variational inference (ELBO) where the prior is approximated from posterior $P(z|x)$.

3 Proposed Approach

We encode each sentence x with a frozen autoencoder (e.g., Bert-GPT2) and consider its sentence representation $E(x)$ as the input of INNs (Figure 2). We propose two training strategies to map the hidden representations into the Gaussian space.

3.1 Training Strategy

Unsupervised INN. Firstly, we train the INN-based model unsupervised, which minimises the negative log-likelihood of the marginal distribution of latent representation $z = E(x)$:

$$\mathcal{L}_{\text{unsup}} = -\mathbb{E}_{x\sim p(x)}\Big[T(E(x))\Big]^2 - \log|T'(E(x))|$$

As the minimisation leads to a bijective mapping between the distributed representation and the disentangled latent representation (multivariate Gaussian space), it allows for a more semantically consistent representation of the geometric (role-content) clustering properties of its latent space, allowing for a more consistent traversal and interpolation (Li et al., 2020b) over the sentence space (Figure 1).

Cluster-supervised INN. According to the findings of (Zhang et al., 2022), the content of the predicate-argument structure/semantic roles can be disentangled over the latent space approximated to multivariate Gaussian learned using the Language VAE setting. Using the same foundation, we train the INN component to learn the embeddings, by minimising the distance between points in the same role-content regions and maximising the distance between points in different regions, based on the explanation embeddings and their corresponding central point from the language autoencoder model. For example, given a sentence "***animals*** *require food for survival*" and its central vector of *ARG0-animals*, the training moves the sentence representation closer to the *ARG0-animals* region centre in the INN latent space. Specifically, during the calculation of the posterior, we replace the mean and variance of the standard Gaussian distribution by the centre point of its cluster and a hyper-parameter, which should be less than one, respectively. In this case, each role-content cluster in the latent space will be mapped to a space where each cluster will have its embeddings more densely and regularly distributed around its centre. The objective function can be described as follows:

$$\mathcal{L}_{\text{sup}} = -\mathbb{E}_{x\sim p_{cluster}(x)}\frac{\Big[T(E(x)) - \mu_{cluster}\Big]^2}{1-\sigma^2} - \log\big|T'(E(x))\big|$$

where $T(E(x))$ learns the transformation from x to $z \sim N(\mu_{cluster}, 1-\sigma^2)$. σ^2 is a parameter which can be empirically determined (in this particular context the optimal value was found to be 0.6). Additional details are provided in Appendix A.

3.2 Geometrical Data Augmentation

Data augmentation, which captures and augments a common or distinct feature across different samples, has been considered a common technique to assist disentanglement, such as in Graph (Li et al., 2021) and Image (Liu et al., 2022) representations, but is still limited in the context of sentence generation. In this work, we consider the vector arithmetic and traversal operators as a systematic mechanism to support data augmentation (via semantically controlled sentence generation) for each role-content cluster, described as follows:

$$(1)\quad \mathrm{v} = average(E'(x_i), E'(x_j))$$
$$(2)\quad \mathrm{v}_{neighbour} = \mathrm{v}[i] \sim N(0,1)_{\forall i\in\{0,..,size(\mathrm{v})\}}$$
$$(3)\quad x_{new} = D'(\mathrm{v}_{neighbour})$$

where $x_k \in S$ (original corpus), E' and D' are the encoder and decoder of Optimus fine-tuned over S. $average$ operation aims to modify the sentence while maintaining the target role-content common

to both x_i and x_j (Zhang et al., 2022). The term $\mathrm{v}[i] \sim N(0,1)$ is introduced to resample each dimension of v in the latent space (i.e., traverse its neighbour) and $x_{new} = D'(\mathrm{v}_{neighbour})$ generates a new sentence. Finally, we only keep the sentences holding the target role-content, where the PArgs/SRs of x are annotated via the *AllenNLP* (Gardner et al., 2018) semantic role labeller. Table 1 lists randomly selected examples from augmented explanations. Full details and the supporting ablation study are provided in Appendices A and D.

Role-content	Augmented sentences
ARG0-animal	an animal requires energy to move an animal requires shelter an animal can use its body to breathe
ARG0-human	humans usually use gasoline humans use coal to make food humans depend on pollinators for survival
PRED-mean	summit mean the top of the mountain colder mean a decrease in heat energy friction mean the product of a physical change

Table 1: Augmented explanations. We also provide more examples in Table 11 for qualitative evaluation.

4 Experiments

For the experiments, we start by focusing on the effect of the supervised INN mechanism to examine its impact on the sentence semantic separability of the distributional latent space defined in Section 2 (detailed in Section 4.1). Next, we examine the localised semantic generation control enabled by such semantic separability via latent interpolation (Section 4.2). Further details of the AutoEncoder model and dataset are provided in Appendix A.

4.1 Disentanglement Encoding Evaluation

We examine the latent space separability (i.e., *natural clustering property* (Bengio, 2013)) of our supervision approach on different predicate-argument/semantic roles. In the context of this work, the thematic roles' labels are not referred to control the generation. Instead, we use the predicate argument position markers, e.g. including *ARG0*, *ARG1*, *PRED(V)*, where each category has *a)* four possible word contents (c_i), or *b)* the same content (i.e., *animal*) with different argument/roles, including *ARG0,1,2*. We provide the reconstructed examples of INNs in Table 24.

Disentanglement between *ARG0* clusters. For *ARG0*, we choose *human*, *animal*, *plant*, and *something* due to having the highest frequency in the original dataset, and evaluate model performance from two directions, including forward and backward mapping. Within forward mapping, we assess the disentanglement of the latent space of the INN model from two perspectives (visualisation and classification metrics). Figure 3 displays the distributions of four role-content clusters over the latent space. As we can observe, after the cluster-supervised training strategy, the embeddings are more concentrated at the center of their cluster, and there is a clear boundary between clusters, indicating a better disentanglement when compared to Optimus and unsupervised INNs.

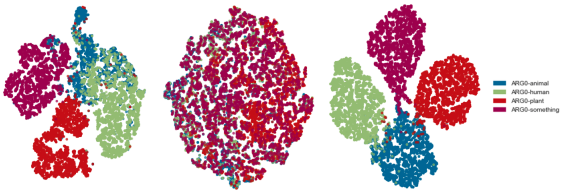

Figure 3: ARG0: t-SNE plot, different colour represents different content regions (blue: animal, green: human, red: plant, purple: something) (left: Optimus, middle: unsupervised, right: cluster supervised), same order for remaining visualizations. We also provide the PCA plot in Figure 11, both visualization shows that supervised embeddings concentrate on the respective cluster center.

We then quantitatively evaluate the disentanglement of ARG0-content clusters. We consider classification task metrics (*accuracy*, *precision*, *recall*, *f1*) as proxies for evaluating region separability, effectively testing cluster membership across different clusters. We choose a non-parametric downstream classifier (i.e., kNN) to quantitatively evaluate the separation of clusters and parametric downstream classifiers, including Naive Bayes (NB) and Support Vector Machine (SVM), to assess both separability and representation capability of latent sentence spaces (Rifai et al., 2011; Conneau et al., 2018). The configuration of the downstream classifiers are detailed in Appendix A.

As shown in table 2, all classifiers trained over supervised latent representations outperformed the unsupervised INN (U) and Optimus (O), indicating that the cluster-supervised approach leads to better disentanglement and representation. Moreover, (O) demonstrates superior performance compared to (U) for the KNN-based evaluation. However, it exhibits lower performance than (U) in NB and SVM. This suggests that the INN-AutoEncoder configuration can more effectively capture sentence semantics (from the point-of-view of AST+distributional content), in the context of a reconstruction task since the VAEs' training process is prone to experiencing posterior collapse.

ARG0: disentanglement proxy metrics					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.972	0.973	0.972	0.972
	U	0.938	0.938	0.938	0.938
	C	**0.979**	**0.979**	**0.979**	**0.979**
NB	O	0.934	0.934	0.933	0.933
	U	0.958	0.958	0.958	0.958
	C	**0.978**	**0.978**	**0.978**	**0.978**
SVM	O	0.970	0.970	0.970	0.970
	U	0.972	0.972	0.972	0.972
	C	**0.980**	**0.980**	**0.980**	**0.980**

Table 2: Disentanglement of ARG0 between Optimus (O), unsupervised INN (U), and cluster-supervised INN (C) where KNN: k-neighbours, NB: naive bayes, SVM: support vector machine. The abbreviations are the same for the remaining tables. Cluster supervision displays consistent improvement with different classifiers.

As for the evaluation of the backward mapping, we calculate the ratio of generated sentences that hold the same role-content as the inputs (henceforth called the invertibility ratio). We randomly selected 100 embeddings as inputs and showed the corresponding ratios in Table 3. We can observe that both unsupervised and supervised cases can achieve high invertibility ratios, indicating that the INN mechanism provides stable invertibility with or without cluster supervision.

ARG0: invertibility ratio (backward: T')				
train	human	animal	plant	something
U	0.980	0.890	0.990	1.000
C	1.000	0.860	0.990	0.950

Table 3: Invertibility test for ARG0, Both INNs with AutoEncoder setup can achieve high ratios, indicating stable invertibility with or without cluster supervision.

Disentanglement between *PRED* clusters. Next, we analyze the disentanglement between *predicate (PRED)* clusters. As shown in Figure 4, although the disentanglement of PRED clusters is not as high as ARG0, the latent space with cluster supervision still performs better than both the unsupervised case and the Optimus model. In Table 4, the supervised INN model achieves better disentanglement, and both unsupervised and supervised could obtain a higher ratio. We also provide the experimental results of *ARG1* disentanglement in Appendix B.

Disentanglement between *ARG0,1,2* clusters. The experiments up to this point investigated the separation between the same pred-argument type but different content clusters. Next, we explore the separability of different pred-argument types with the same content. We thus focus on the *animal* cluster, and investigate the disentanglement between

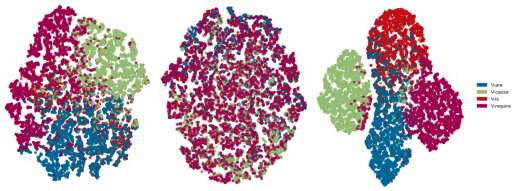

Figure 4: PRED: t-SNE plot (blue: are, green: cause, red: is, purple: require). PCA plot is in Figure 13.

PRED: disentanglement proxy metrics (forward: T)					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.911	0.914	0.910	0.911
	U	0.869	0.873	0.865	0.868
	C	**0.922**	**0.927**	**0.918**	**0.922**
NB	O	0.865	0.866	0.866	0.865
	U	0.873	0.874	0.871	0.872
	C	**0.903**	**0.903**	**0.902**	**0.903**
SVM	O	0.902	0.902	0.903	0.902
	U	0.905	0.906	0.902	0.904
	C	**0.910**	**0.912**	**0.909**	**0.910**

Table 4: Forward evaluation for predicate clusters, the invertibility ratio and statistical significance test are provided in Table 14 and 18.

ARG0-animal, *ARG1-animal*, and *ARG2-animal*. As illustrated in Figure 5, the animal clusters with different pred-argument types can be separated after cluster-supervised training, which indicates that the INN model can capture the difference between the same content with different pred-argument type in the case of similar topic, indicating the INN-based approach could jointly learn separable embeddings w.r.t. role-content and content alone.

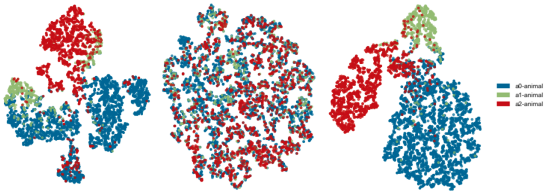

Figure 5: Animal: t-SNE plot (blue: ARG0-animal, green: ARG1-animal, red: ARG2-animal), PCA plot is provided in Figure 14.

Table 5 and 15 show the disentanglement metrics and the invertibility ratio, respectively. Similarly to the previous experiment, the supervised case outperforms both the unsupervised and the Optimus models. Both INNs can achieve an invertibility ratio of at least 90%.

4.2 Disentanglement Decoding Evaluation

Finally, we evaluate the disentangled sentence geometry from the perspective of sentence generation. We specifically focus on linear interpolation as it can provide more efficient traversal between sen-

Animal: disentanglement metrics (*f1 score*)			
train	KNN	NB	SVM
O	0.960	0.928	0.946
U	0.958	0.930	0.947
C	**0.967**	**0.937**	**0.950**

Table 5: Forward evaluation for Animal, we only show *f1* since the same value across different metrics. Results indicate improved separation across role clusters.

tences and clusters than other traversal approaches (e.g., *Ornstein-Uhlenbeck*), commonly used in the NLP domain (Li et al., 2020b) and in the evaluation of disentanglement (Bengio, 2013).

Interpolation localisation. Firstly, we evaluate the localisation of latent interpolation that interpolates a path $z_t = z_1 \cdot (1-t) + z_2 \cdot t$ with t increased from 0 to 1 by a step size of 0.1, where z_1 and z_2 represent the latent representations of source and target sentences. As a result, 9 sentences are generated on each interpolation step. On a latent space with better token-level role-content separation, given two sentences with the same role-content as endpoints, we can observe that the intermediate sentence can hold the same role-content during interpolation.

In terms of qualitative evaluation, Table 6 provides the interpolation paths of cluster-supervised INN and Optimus, as for Optimus, we can observe that the intermediate explanations could transition smoothly from source to target for *argument*. However, the *predicate* is more abruptly changed, indicating lower *predicate-content* disentanglement (e.g., *predicate-require* and *predicate-eat*). Instead, the supervised INN can fix the *predicate-require* during interpolation, indicating better separability between different predicate-content results in better generation control. More examples are provided in Table 22 and 23. We then quantitatively evaluate the localisation of interpolation. We randomly select 200 sentence pairs from the dataset holding the same role-content and report the ratio of intermediate sentences with the same role-content as inputs. As illustrated in Figure 6, the intermediate sentences from the supervised INN can better hold the same role-content as inputs, especially for *predicate* which usually has a lower effect on distributional sentence semantics (Zhang et al., 2022), indicating that our supervision can lead to better latent space separability and generation control.

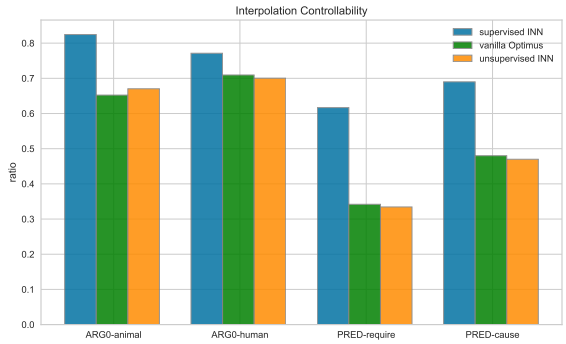

Figure 6: Interpolation control evaluation, we can observe that supervised INN with better semantic separability can lead to better localised semantic control.

interpolation localisation: *predicate-require*

source: humans require freshwater for survival

Optimus:
1. humans require water and food through fossil fuels
2. humans require water for survival
3. humans produce small amounts of consumer food
4. human has a positive impact on a plant's survival
5. humans convert food into animal prey
6. humans make food for themselves by eating
7. animals require food for survival
8. animals require nutrients from the air
9. humans eat plants for food
10. animals require food for survival

Cluster-supervised INN:
1. humans require water for survival
2. nonhumans require water for survival
3. animals require water and food
4. animals require water to survive
5. animals require water to live
6. animals require food for survival
7. animals require food for survival
8. animals require food for survival
9. animals require food for survival
10. animals require food to survive

target: animals require food to survive

Table 6: Interpolation examples, indicating the cluster-supervised INN can provide better localised/symbolic semantic control. We also report the interpolations of AutoEncoder and unsupervised INN in Table 21.

Interpolation smoothness. Moreover, we quantitatively evaluate the latent space geometry via interpolation smoothness metrics (IS, Zhang et al. (2024)), which calculates the ratio between the ideal semantic distance (i.e., the aligned semantic distance between source and target sentences) and the actual semantic distance (i.e., the sum of semantic distance between adjacent sentences during interpolation). A higher ratio indicates that the actual path aligns better with the ideal path, suggesting better semantic-geometric properties. The metric is defined as: IS $= \mathbb{E}_{(s_0,\ldots,s_T)\sim P} \frac{\delta(\text{align}(s_0,s_T))}{\sum_{t=0}^{T} \delta(\text{align}(s_t,s_{t+0.1}))}$ where $s_0, \ldots, s_T$ is the sequence of sentences during interpolation, δ and align are sentence similarity and alignment

Evaluation Metrics	avg IS↑	max IS↑	min IS↑
DAE (Vincent et al., 2008)	0.144	0.330	0.055
AAE (Makhzani et al., 2015)	0.142	0.284	0.054
LAAE(Rubenstein et al., 2018)	0.172	0.347	0.056
DAAE (Shen et al., 2020)	0.055	0.061	0.023
β-VAE (Higgins et al., 2016)	0.198	0.379	0.041
AdaVAE (Tu et al., 2022)	0.085	0.105	0.050
Della (Hu et al., 2022)	0.253	0.416	0.155
Optimus (Li et al., 2020b)	0.220	0.525	0.130
AutoEncoder (Bert-GPT2)	0.259	0.585	0.165
INN (U) (our)	0.251	0.540	0.159
INN (C) (our)	**0.282**	**0.607**	**0.206**

Table 7: Geometrical examination via IS metric.

functions, respectively, which are performed via Word Mover's Distance (Zhao et al., 2019). We choose the standard language VAE baselines (i.e., the prior is the std. Gaussian distribution). Their implementation details are provided in Appendix A. We randomly sample 200 sentence pairs and report the IS metric. As illustrated in Table 7, our model can deliver smoother interpolations comparatively to the baselines, indicating semantic disentanglement can lead to better latent space geometry.

5 Related Work

Sentence representation. Sentence representations are usually trained in supervised (Conneau et al., 2017; Reimers and Gurevych, 2019), constrastive (Giorgi et al., 2021; Yan et al., 2021; Chuang et al., 2022), or generation-oriented (Wang et al., 2021; Wu and Zhao, 2022; Chuang et al., 2022) fashion. Recent work (Huang et al., 2023) explored the compositional sentence representation for improved explainability and generation. However, these works still lack the emphasis on the geometric interpretation and control of the underlying sentence space, which this work focused on.

Sentence disentanglement. In the Natural Language Generation domain, most previous investigations explored the disentanglement between two specific linguistic perspectives, such as sentiment-content (John et al., 2019; Li et al., 2022), semantic-syntax (Bao et al., 2019), and negation-uncertainty (Vasilakes et al., 2022), or syntactic disentanglement (Mercatali and Freitas, 2021; Felhi et al., 2022). In this work, we provide a formal-geometrical lens, with the support of *argument structures* as a sentence representation model, for sentence disentanglement targeting for localised semantic control. This work is the first integration of flow-based INN mechanisms to improve disentanglement, separation and semantic control of sentence spaces.

INNs in NLP. Şahin and Gurevych (2020) concentrate on modelling morphological inflection and lemmatization tasks, utilizing INN to learn a bijective transformation between the word surface and its morphemes. Li et al. (2020a) proposed BERT-flow, transforming sentences from a BERT sentence space to a standard Gaussian space. Ding and Gimpel (2021) deployed flow-based INN to enrich VAE prior distribution, while Gu et al. (2023) use flow mechanisms to control attributes in style transfer tasks. This work focused on semantic separability, geometrical operations and control over the distributed representation of sentences. Moreover, this work is the first to explore geometrical data augmentation to support semantic disentanglement.

6 Conclusions and Future Work

This work focused on an INN-based mechanism to support better disentangled and separated latent sentence spaces over language autoencoders. By aligning the predicate-argument structure of sentences to the latent representations, we aimed to build a bridge between the formal and distributional semantics perspectives for sentence representation. We define the sentence semantic disentanglement from the perspective of formal semantics, aligning the predicate-argument structure to disentanglement and cluster separation properties, and exploiting the invertibility and bijection properties of INNs to facilitate such alignment. Experimental results indicate that the invertibility mechanisms can transform the distributed hidden space of an autoencoder into a latent space where AST-level syntactic and semantic transformations can be localised, interpolated and controlled. Secondly, we propose a supervision approach, which leads to an improved disentangled and separated space. This property can facilitate localised interpolation control. Lastly, we utilise these geometric properties to support a semantically controlled data augmentation to assist the disentanglement process.

Since our work connects distributional and formal semantics via disentanglement, one future direction is to explore the safety and control of the formal semantic properties of Large Language Models. Besides, recent work (Liu et al., 2023) revealed that distinct factors can be composed by modelling the moving of latent vectors via ordinary differential equations, which can be adapted to sentence representations to deliver more complex sentence transformations within the latent space.

7 Limitations

This work focused on the disentangled sentence representations geometry to deliver localised/semantic/formal semantic control. While this work is motivated by providing more localised distributed representations, which can positively impact the safety and coherence of generative models, few scoping observations need to be established:

1. The specific safety guarantees of these models are not fully established.

2. In the Image domain, the generative factors of disentangled latent image spaces can be controlled by simply resampling (e.g., Ornstein-Uhlenbeck random walk) over the corresponding dimensions (Higgins et al., 2017; Kim and Mnih, 2018) or by traversing within geometrical boundaries (Jeon et al., 2019). Those techniques cannot be directly adopted in latent sentence spaces.

3. While the language autoencoder with unsupervised INN exhibit a distinct learning pattern with regard to semantic distribution, further understanding is required in terms of information bottleneck properties (Saxe et al., 2018) and on the semantic distribution of unsupervised INNs in language modelling tasks.

4. Furthermore, this study exclusively focused on a corpus of sentences which are conceptually dense ((Dalvi et al., 2021)). The exploration of its performance on other types of sentences, including sentences with complex clausal-phrasal constructions, or sentences with non-compositional idioms, is yet to be undertaken.

References

Lynton Ardizzone, Till Bungert, Felix Draxler, Ullrich Köthe, Jakob Kruse, Robert Schmier, and Peter Sorrenson. 2018-2022. Framework for Easily Invertible Architectures (FrEIA).

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.

Benjamin Bengfort and Rebecca Bilbro. 2019. Yellowbrick: Visualizing the Scikit-Learn Model Selection Process. 4(35).

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

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. 2018. Isolating sources of disentanglement in vaes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2615–2625.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

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

Xiaoan Ding and Kevin Gimpel. 2021. FlowPrior: Learning expressive priors for latent variable sentence models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3242–3258, Online. Association for Computational Linguistics.

Laurent Dinh, David Krueger, and Yoshua Bengio. 2014. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

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.

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

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

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.

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

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.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

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.

James Y. Huang, Wenlin Yao, Kaiqiang Song, Hongming Zhang, Muhao Chen, and Dong Yu. 2023. Bridging continuous and discrete spaces: Interpretable sentence representation learning via compositional operations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14584–14595, Singapore. Association for Computational Linguistics.

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

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

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018a. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC*

2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Peter A Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. 2018b. 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.

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.

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.

Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

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

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.

Haoyang Li, Xin Wang, Ziwei Zhang, Zehuan Yuan, Hang Li, and Wenwu Zhu. 2021. Disentangled contrastive learning on graphs. *Advances in Neural Information Processing Systems*, 34:21872–21884.

Zhuang Li, Lizhen Qu, Qiongkai Xu, Tongtong Wu, Tianyang Zhan, and Gholamreza Haffari. 2022. 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.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2023. 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.

Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q O'Neil, and Sotirios A Tsaftaris. 2022. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Gary F Marcus. 2003. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

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.

Nathan Michlo, Richard Klein, and Steven James. 2023. Overlooked implications of the reconstruction loss for vae disentanglement. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4073–4081.

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

Ryan M. Nefdt. 2020. A puzzle concerning compositionality in machines. *Minds and Machines*, 30(1):47–75.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Neural modeling of multi-predicate interactions for Japanese predicate argument structure analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1591–1600, Vancouver, Canada. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Karl Ridgeway and Michael C Mozer. 2018. Learning deep disentangled embeddings with the f-statistic loss. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 185–194.

Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. 2011. The manifold tangent classifier. In *Neural Information Processing Systems*.

Paul K Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. 2018. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*.

Gözde Gül Şahin and Iryna Gurevych. 2020. Two birds with one stone: Investigating invertible neural networks for inverse problems in morphology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7814–7821.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. 2018. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*.

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.

Jonathon Shlens. 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

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

Haoqin Tu, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. 2022. Adavae: Exploring adaptive gpt-2s in variational auto-encoders for language modeling. *arXiv preprint arXiv:2205.05862*.

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.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA. Association for Computing Machinery.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*.

Bohong Wu and Hai Zhao. 2022. Sentence representation learning with generative objective rather than contrastive objective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3356–3368, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuangzhi Wu, Dongdong Zhang, Nan Yang, Mu Li, and Ming Zhou. 2017. Sequence-to-dependency neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–707, Vancouver, Canada. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Sygns: A systematic generalization testbed based on natural language semantics. *arXiv preprint arXiv:2106.01077*.

Yingji Zhang, Danilo S Carvalho, Ian Pratt-Hartmann, and André Freitas. 2022. Quasi-symbolic explanatory nli via disentanglement: A geometrical examination. *arXiv preprint arXiv:2210.06230*.

Yingji Zhang, Danilo S. Carvalho, Marco Valentino, Ian Pratt-Hartmann, and Andre Freitas. 2024. Improving semantic control in discrete latent spaces with transformer quantized variational autoencoders.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. 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. Table 9 illustrates the semantic, structure, and topic information of explanatory sentences over the latent space.

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

Table 8: Statistics from explanations datasets.

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. More details about semantic structure and lexical information are provided in Table 9 and 10.

Data Augmentation. Algorithm 1 illustrates the detailed process of data augmentation. The key aspect of data augmentation is to keep the data distribution unchanged while increasing the size of the dataset. Therefore, during traversal, we only sample the value whose probability density is between 0.495 and 0.505. In other words, for each original explanation, we only traverse its close neighbours over the latent space. We increased the number of explanations in each role-content cluster to 3000 and kept the balance of each role-content category. We provide more qualitative examples in Table 11. Moreover, we visualise latent semantic distribution before and after augmentation in Figure 7. As we can observe, the data augmentation can maintain the semantic distribution unchanged. For example, *PRED-is* (red colour in the right column) is widely distributed over the latent space before and after augmentation. *ARG0-something* (purple colour in the left column) is far from other clusters with or without data augmentation in latent space.

Downstream Classifier. In this experiment, we apply three downstream classifiers, including non-parametric classifier: k-nearest neighbours (KNN) and parametric classifiers: Naive Bayes (NB) and Support Vector Machine (SVM), to evaluate the

Algorithm 1 Data Augmentation

Define: R as the role set (ARG0, PRED, ...).
Define: C as the content set (vocabulary).
Define: S as the explanation corpus (sentences).
Define: $s = [(c_1, r_1), ..., (c_i, r_i)] \in S,\ c_i \in C,\ r_i \in R$ as a sentence.
Define: $(c_t, r_t) \mid r_t \in R,\ c_t \in C$ as the target role-content (e.g., ARG1-animal).
Define: $S_t = \forall s \in S \mid \exists (c_k, r_k) = (c_t, r_t)$ as the set of sentences with the target role-content.
Define: $E(s) : S \rightarrow \mathbb{R}^n$ as encoder (embedding) function.
Define: $D(vec) : \mathbb{R}^n \rightarrow S$ as the explanation decoded from Decoder D.
Define: L: list for keeping augmented sentences.
Define: $SRLer(s)$: semantic role label annotator for s.
for all $(s_i, s_j) \in S_t,\ s_i \neq s_j$ **do**
  $vec = average(E(s_i), E(s_j))$
  for all $vec[i] \in vec$ **do**
    $vec[i] = N(0, 1)$ # *neighbour traversal*
    $s_n = D(vec)$ # *new sentence*
    if $s_n \notin L$ **AND** $R \in SRLer(s_n)$ **then**
      put s_n in L.
    end if
  end for
end for

Cluster	Theme, Pattern, and Explanatory sentences
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: Semantic, structure, topic information of explanatory sentences, where the cluster is the categories of k-means classifier.

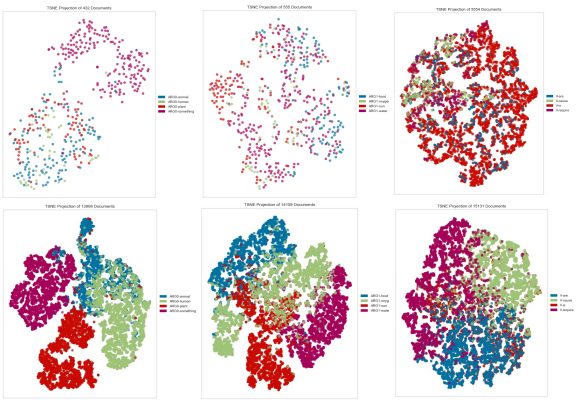

Figure 7: t-SNE plot for Data augmentation (top: original dataset distribution, bottom: augmented dataset distribution), (left: ARG0-animal(blue), human(green), plant(red), something(purple); middle: ARG1-food(blue), oxygen(green), sun(red), water(purple); right: PRED-are(blue), cause(green), is(red), require(purple).

separability of latent representation. Those classifiers and classification metrics are implemented based on *scikit-learn* package (Pedregosa et al., 2011) with default hyper-parameters. We train those classifiers on the training set ($\times 60\%$) and evaluate them on the test set ($\times 40\%$). For multi-class classification, we set *macro* for *precision*, *recall*, and *f1* since *macro*-averaged metric for each class is calculated independently, and then the average is taken, which ensures that the performance of the model in each class contributes equally to the final metric, regardless of the class size.

Visualizer. In this experiment, we implement t-SNE and PCA visualisation based on *Yellowbrick* library (Bengfort and Bilbro, 2019)[3]. We empirically set $decompose_by = 4$ for all cases. However, we found no significant difference between different $decompose_by$ parameters.

Baselines for Interpolation Smoothness. In the experiment, we implement five LSTM-based autoencoders, including denoising AE (Vincent et al. (2008), DAE), β-VAE (Higgins et al., 2016), adversarial AE (Makhzani et al. (2015), AAE), label adversarial AE (Rubenstein et al. (2018), LAAE), and denoising adversarial autoencoder (Shen et al. (2020), DAAE). Their implementation relies on the open-source codebase available at the URL [4]. As for transformer-based VAEs, we implement Optimus (Li et al., 2020b), AdaVAE (Tu et al., 2022)[5], and Della (Hu et al., 2022)[6]. All baseline models undergo training and evaluation with the hyper-parameters provided by their respective sources. A latent dimension of 32 is specified to ensure a

[3] https://www.scikit-yb.org/en/latest/api/text/tsne.html

[4] https://github.com/shentianxiao/text-autoencoders

[5] https://github.com/ImKeTT/AdaVAE

[6] https://github.com/OpenVLG/DELLA

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 appear in explanations corpus.

Role-content	Augmented sentences
ARG0-plant	plants use sunlight often to make food for themselves plants produce light in the winter by photosynthesizing green plants contain (water ; food) plants take in oxygen from the air a plant requires water in order to perform photosynthesis some plants grow organically plants use soil as a source of water
ARG1-water	water is liquid by volume salt water is a kind of solution water is two things together water is boiling in the pot water is an (inexhaustible ; wasteable) resource water is an (electrical ; electrical energy) insulator water is a part of soup
ARG2-animal	a hurricane is a kind of animal a bird is a kind of animal a sperm whale is a kind of animal a wren is a kind of animal a dog is a kind of native animal a chameleon is a kind of animal
PRED-require	making tools requires using sharp tools plants require resources to provide food for themselves a system requires electrical energy to operate crops require specialized environments to grow cooking requires food from human food chain producing an object requires chemical energy living things require energy from the sun for survival growth requires the production of more cells

Table 11: Qualitative evaluation of geometrical data augmentation.

uniform and equitable comparative analysis.

Autoencoder. In this work, we employ an autoencoder architecture with the same configuration as described in (Li et al., 2020b)[7]. The encoder component is based on BERT (Devlin et al., 2018), while the decoder component is based on GPT2 (Radford et al., 2019). The latent space dimension is set to 32 (low-dimension) as Michlo et al. (2023) revealed that strong compression, such as strong KL regularisation term in ELBO, can lead to the phenomenon of disentanglement of images.

To establish the connection between the encoder and decoder, the input sentence x is first encoded by BERT into the latent space, denoted as $N(\mu, \Sigma)$. The parameters μ and Σ are trainable and determine the mean and covariance of the Gaussian distribution. Next, a sample $z \sim N(\mu, \Sigma)$ is passed through a multi-layer perceptron called W. This step expands the dimensionality of z to obtain a fixed-length embedding $h \in \mathbb{R}^{D \times L \times H}$, where D represents the dimensions of the heads, L is the number of heads, and H is the number of hidden layers. The latent space injection can be described as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{Q[z; K]^T}{\sqrt{d}})[z; V]$$

[7] https://github.com/ChunyuanLI/Optimus

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

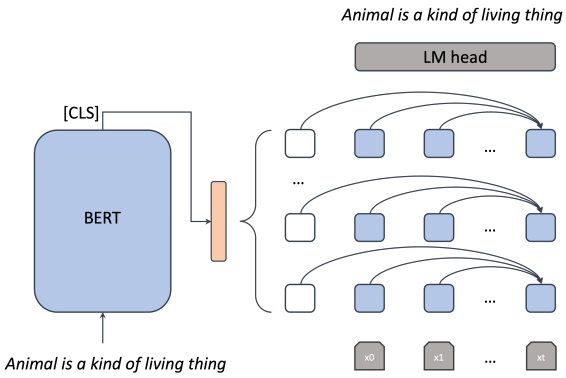

Figure 8: Latent sentence injection.

INN. The INN consists of 10 invertible blocks. Each is built from three layers, including an affine coupling (Dinh et al., 2016), permutation layer, and ActNorm (Kingma and Dhariwal, 2018). Figure 9 displays one single invertible block. The model was implemented using the FrEIA library (Ardizzone et al., 2018-2022) [8]. As for training hyperparameters of INN, firstly, both input and output have the same dimensions as the latent space dimension of the autoencoder. Secondly, inside the affine coupling block, the sub-network is MLP with 512 as the hidden dimension. Thirdly, we use AdamW (Loshchilov and Hutter, 2017) to optimise the model where the learning rate is 5e-04 in the experiment.

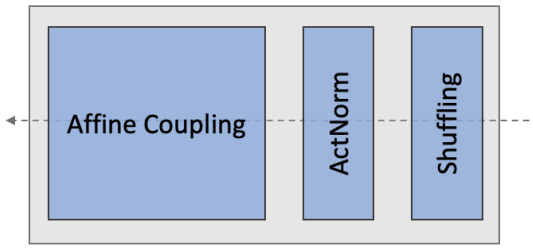

Figure 9: INN one single block.

The forward process of the affine coupling layer can be described as follows:

$$\begin{aligned} x_a, x_b &= \text{split}(x) \\ \log s, t &= m_\theta(x_b) \\ s &= \exp(\log s) \\ y_a &= s \odot x_a + t \\ y_b &= x_b \\ y &= \text{concat}(y_a, y_b) \end{aligned} \tag{1}$$

[8] https://github.com/VLL-HD/FrEIA

Where m_θ is a two-layer neural network. x and y are the input and output. The reversed process is:

$$\begin{aligned} y_a, y_b &= \text{split}(y) \\ \log s, t &= m_\theta(y_b) \\ s &= \exp(\log s) \\ x_a &= (y_a - t)/s \\ x_b &= y_b \\ y &= \text{concat}(x_a, x_b) \end{aligned} \tag{2}$$

B Additional Supervision Results

Disentanglement between *ARG1* clusters We consider four *ARG1* clusters, including *ARG1-food*, *ARG1-oxygen*, *ARG1-sun*, *ARG1-water*, and evaluate model performance following the same procedure. Figure 10 displays the distributions of four role-content clusters over the latent space. With similar observations as before, the INN cluster-supervised training strategy can learn better disentanglement between ARG1 clusters. Table 12

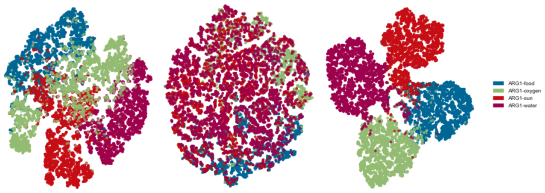

Figure 10: ARG1: t-SNE plot (blue: *food*, green: *oxygen*, red: *sun*, purple: *water*). Supervision (right) induces separability comparable with ARG0. PCA plot is provided in Figure 12.

and 13 show the disentanglement metrics and invertibility ratio, respectively. With similar observations as the previous experiment: all classifiers trained over the supervised latent representation outperform both the unsupervised INN model and Optimus, and both unsupervised and supervised cases can achieve higher ratios (at least 0.95).

Invertibility ratio. Table 13, 14, and 15 report the invertibility test for *ARG1*, *PRED*, and *ARG0,1,2* clusters, respectively. We can observe that INN with both training approaches can perform stable invertibility.

Traversal decoding for *Animal* clusters. Table 16 shows the decoded explanations traversed around the central point of each cluster in the latent space of cluster-supervised INN.

Traversal decoding for cluster connection. Table 17 displays the decoded middle points between clusters. It is also observable that there are low-density embedding regions at the transition (con-

ARG1: disentanglement proxy metrics (forward: T)					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.934	0.934	0.933	0.933
	U	0.914	0.914	0.914	0.913
	C	**0.954**	**0.954**	**0.954**	**0.954**
NB	O	0.904	0.910	0.902	0.904
	U	0.922	0.922	0.922	0.922
	C	**0.957**	**0.957**	**0.957**	**0.957**
SVM	O	0.951	0.951	0.951	0.950
	U	0.953	0.953	0.952	0.953
	C	**0.959**	**0.959**	**0.959**	**0.959**

Table 12: Forward evaluation for ARG1, consistent results on different classifiers indicate that supervision can perform better semantic disentanglement.

ARG1: invertibility ratio (backward: T')				
train	food	oxygen	sun	water
U	0.990	0.980	0.950	1.000
C	0.960	0.950	0.960	1.000

Table 13: backward evaluation for ARG1 clusters. unsupervised INN (U), and supervised INN (S).

nection) between two clusters. We decode the middle datapoints between *animal* and *human* clusters and list them in Table 17. From those examples, we can observe that such explanations are related to both *animal* and *human*. This result implies that the explanations may be geometrically represented in a similar way as they were originally designed in the WorldTree corpus (maximising lexical overlaps for pred-arg alignments within an explanation chain) for supporting multi-hop inference tasks.

Principal component analysis (PCA) visualisation. In addition to the non-linearised t-SNE plot, we also provide linearised visualisation via PCA (Shlens, 2014). Figure 11,12,13, and 14 visualize the separation of *ARG0*, *ARG1*, *PRED*, and *animal*. Similar to the observation before, cluster supervision can lead to better separation and cluster.

Figure 11: PCA visualization for *ARG0*.

C Statistical Significance Tests for *PRED* Downstream Classifiers

Statistical significance testing is a standard statistical tool devised to ensure that experimental results are not coincidental and reliable. Following

PRED: invertibility test (backward: T')				
train	is	are	cause	require
U	1.000	0.950	0.970	0.800
C	1.000	0.880	0.900	0.820

Table 14: backward evaluation for predicate clusters. unsupervised INN (U), and supervised INN (S).

Animal: invertibility ratio (backward: T')			
train	ARG0	ARG1	ARG2
U	0.990	0.990	0.900
C	0.970	0.960	0.920

Table 15: Backward evaluation for Animal.

Traversing Animal clusters

1: animals must escape from predators
2: animals require air to breathe
3: an animal requires warmth for survival

1: animals are small in size
2: animals usually are not carnivores
3: animals are a part of an environment

1: a rabbit is a kind of animal
2: an otter is a kind of animal
3: a horse is a kind of animal

Table 16: Traversal in each cluster (top: *ARG0-Animal*, middle: *ARG1-Animal*, bottom: *ARG2-Animal*).

Cluster connection

1. humans sometimes hunt animals that are covered in fur
2. animals / human habitats require food
3. an animal may be bred with a human for food
4. animals eat humans
5. a human can not eat algae and other animals

Table 17: Middle explanations between *ARG0-animal* and *ARG0-human*.

Figure 12: PCA visualization for *ARG1*.

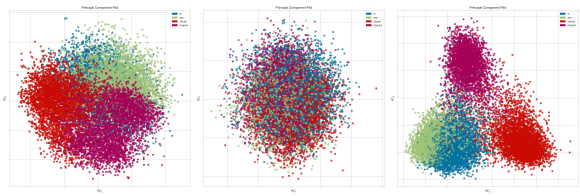

Figure 13: PCA visualization for *PRED*.

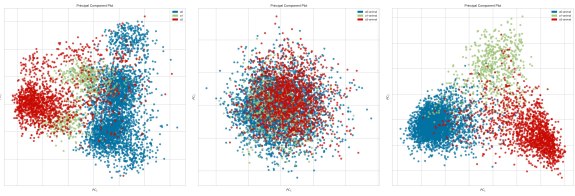

Figure 14: PCA visualization for *Animal*.

the work (Dror et al., 2018)[9], we provide statistical significance tests to rigorous and quantitatively evaluate the stability of trainable downstream classifiers, which indirectly indicates the representation capability.

Our attention was directed towards *PRED* clusters due to the comparatively decreased performance of downstream classifiers within this category as *PRED* usually contains less semantic information (Zhang et al., 2022). We select *accuracy* metric, set $\alpha = 0.05$, and choose *bootstrap* statistical test which was used with a variety of NLP tasks (Ouchi et al., 2017; Wu et al., 2017).

As illustrated in Table 18, (1) the U-C pair consistently yields a diminished significance value, suggesting reliable classification performance resulting from superior representational capabilities facilitated by the AutoEncoder with INN configuration, compared with Optimus. (2) the scores of (O-C) pairs are consistently lower than those of (O-U) pairs, indicating our supervision (C) can better represent semantic information than unsupervised INN. We refer (Dror et al., 2018) for an in-depth illustration of statistical significance tests in NLP.

Statistical significance tests for *PRED*		
classifier	source	Bootstrap (p-value)↓
KNN	O-C	0.0155
	U-C	**0.0000**
	O-U	1.0000
NB	O-C	0.0000
	U-C	**0.0000**
	O-U	0.2268
SVM	O-C	0.3594
	U-C	**0.0000**
	O-U	1.0000

Table 18: Statistical significance tests for downstream classifiers (O: Optimus, U: unsupervised INN, and C: cluster supervised INN). We **highlight** the best significant test value, indicating reliable classification performance derived from better representation capability.

D Ablation of Data Augmentation

***PRED* semantic role.** Firstly, we analyse the effect of our supervision approach on *PRED* semantic role with three lexical contents without data augmentation, including *are* (×449), *cause* (×380), and *require* (×262). The rationale for their selection is that they are less frequent in corpus and partially overlap in latent space. Moreover, the contents under *PRED* usually have less effect on the contextual semantics (Zhang et al., 2022). Those difficulties allow us to fairly analyse the effect of our supervision approach. Following a similar order, we first visualise the t-SNE and PCA plots in Figure 15. As we can observe, the cluster-supervised approach can better represent the cluster and separation for different contents under *PRED* semantic role label without data augmentation. Next, we apply downstream classifiers to evaluate cluster separation. As illustrated in Table 19, our cluster-supervised approach results in better classification performance, indicating better disentanglement.

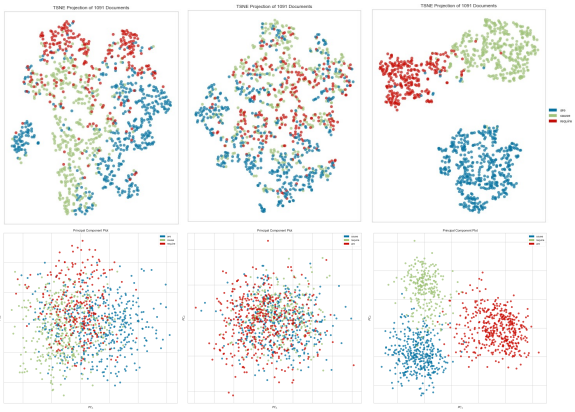

Figure 15: Ablation: t-SNE plot (top), PCA plot (bottom) (left: Optimus, middle: unsupervised, right: cluster-supervised) where blue: *PRED-are*, green: *PRED-cause*, red: *PRED-require*.

PRED: disentanglement proxy metrics					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.858	0.847	0.844	0.846
	U	0.837	0.849	0.827	0.830
	C	**0.965**	**0.963**	**0.961**	**0.962**
NB	O	0.839	0.823	0.833	0.826
	U	0.901	0.895	0.891	0.893
	C	**0.977**	**0.974**	**0.975**	**0.974**
SVM	O	0.876	0.863	0.866	0.865
	U	0.954	0.953	0.949	0.950
	C	**0.967**	**0.965**	**0.967**	**0.966**

Table 19: Ablation: disentanglement proxy metrics for *PRED-are*, *PRED-cause*, and *PRED-require*.

***ARG0* semantic role.** Next, we provide the same analyse for fewer frequent *ARG0* clusters: *ARG0-animal* (×126), *ARG0-human* (×43), *ARG0-plant*

[9] https://github.com/rtmdrr/testSignificanceNLP/tree/master

(×77), and *ARG0-something* (×186). As illustrated in Figure 16, cluster supervision can lead to better role-content separation/disentanglement. Moreover, we can observe that cluster-supervision leads to better proxy disentanglement metrics in Table 20.

Furthermore, compared with Table 2, the incorporation of latent representation with data augmentation results in enhanced classification performance. This observation implies that our data augmentation technique can more effectively capture semantic information, thereby aiding downstream classifiers.

Figure 16: Ablation: t-SNE plot (top), PCA plot (bottom) (left: Optimus, middle: unsupervised, right: cluster-supervised) where blue: *ARG0-animal*, green: *ARG0-human*, red: *ARG0-plant*, purple: *ARG0-something*.

ARG0: disentanglement proxy metrics					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.890	0.890	0.850	0.867
	U	0.890	0.896	0.834	0.858
	C	**0.919**	**0.907**	**0.858**	**0.877**
NB	O	0.855	0.809	0.784	0.792
	U	0.936	0.916	0.905	0.910
	C	**0.965**	**0.958**	**0.950**	**0.954**
SVM	O	0.843	0.630	0.691	0.656
	U	0.895	0.847	0.770	0.782
	C	**0.901**	**0.935**	**0.779**	**0.790**

Table 20: Ablation: disentanglement proxy metrics for *ARG0-animal*, *ARG0-human*, *ARG0-plant*, and *ARG0-something*.

E Controlled Interpolation

In tables 22 and 23, we provide more controllable interpolation examples. Those examples reveal that the latent space with better role-content separation from supervised INN can provide better interpolation control, indicating better latent space geometry.

Interpolation localisation: *predicate-require*

source: humans require freshwater for survival

AutoEncoder:
1. humans require water to survive
2. marine mammals require great amounts of water
3. animals require oxygen to survive
4. animals require water for survival
5. animals must eat water to survive
6. animals require water and food
7. animals require water for survival
8. animals must eat to survive
9. animals require food for survival
10. animals must eat food to survive

Unsupervised INN:
1. nonhumans require water to survive
2. marine animals require food for survival
3. animals must breath to survive
4. animals require water for survival
5. animals require water from their ecosystems
6. animals require water for survival
7. animals must eat food for survival
8. animals require food for survival
9. animals require food for survival
10. animals require food for survival

target: animals require food to survive

Table 21: Interpolation examples where top and bottom sentences are source and target, respectively.

F INNs: Explanation Reconstruction

Table 24 shows some reconstructed explanations from AutoEncoder, unsupervised INN, and supervised INN, respectively.

Interpolation localisation: *predicate-is*

source: the sun is in the northern hemisphere

1. the sun is located in the northern hemisphere
2. the sun is in the northern hemisphere
3. the sun is made of air around the sun
4. the sun is a source of sunlight for organisms
5. the sun is a source of sunlight for birds
6. the sun is a source of energy for organisms living in an arctic environment
7. the sun is a source of food for plants
8. food is a source of oxygen ; water for plants
9. food is a source of energy for plants by producing heat
10. food is a source of energy for a plant or animal / living thing

1. the sun is the dominant star in the night sky
2. the sun is closer to the earth than it is to the sun
3. the sun is a star in the night sky
4. the sun is good for the environment by providing sunlight to plants
5. the atmosphere is an environment for intensive farming
6. the respiratory system carries oxygen to the rest of the body
7. food contains nutrients ; water ; food energy
8. food is the nutrient for (plants ; animals)
9. producers are a source of energy for producers by weathering
10. food is a part of a plant / animals / living things

target: food is a source of energy for animals / plants

Table 22: Interpolation examples (top: supervised INN, bottom: Optimus).

Interpolation localisation: *argument-animals* and *predicate-require*

source: animals require food to survive

1. animals require water to survive
2. animals require food for survival
3. animals require food for survival
4. animals require nutrients from food
5. an animal requires food for survival
6. an animal requires food for survival
7. an animal requires nutrients from producers
8. an animal requires nutrients for survival
9. an animal requires nutrients from food
10. an animal requires nutrients from producers

1. animals need sunglasses for protection
2. animals live in an environment
3. animals need food to thrive
4. animals require energy for survival
5. a consumer uses some of the food that is available
6. only a producer eats plants
7. a human produces its own food
8. an animal requires nutrients in a source of food to survive
9. an animal requires energy to perform photosynthesis
10. an animal requires nutrients to grow

target: an animal requires nutrients from producers

Table 23: Interpolation examples (top: supervised INN, bottom: Optimus).

Augmented explanations	BERT-GPT2	unsupervised INN	supervised INN
a animal requires water for survival	a animal requires water for survival	a animal requires water for survival	a animal requires water for survival
an animal requires a mate for survival	an animal requires a mate to reproduce	an animal requires a mate to reproduce	an animal requires a reproductive system for survival
some animals sometimes hunt for prey	some animals prey on other animals	some animals sometimes catch prey	some animals sometimes hunt for prey
an animal requires energy of its own to move	an animal requires energy from somewhere to move	an animal requires energy to move	an animal requires energy for movement
an animal requires energy to run	an animal requires energy to run	an animal requires energy to run	an animal requires energy to run
animals live in their habitats	animals live in their habitats	animals live in their habitat	animals live in their habitat
animals must eat animals to survive	animals must eat to survive	animals must eat other animals to survive	animals must eat to survive
animals taste flavors	animals taste flavors	animals taste flavors	animals taste flavors
animals eat plants	animals eat plants	animals eat plants	animals eat plants
an animal requires nutrients to grow and heal	an animal requires nutrients in soil for survival	an animal requires nutrients to grow and repair	an animal needs to store fat to grow
animals require oxygen to grow	animals require oxygen to grow	animals require oxygen to breath	animals require oxygen for survival
an animal needs to breathe in order to survive	an animal requires food for survival	a animal needs to breathe to survive	an animal requires water and food to survive
humans cause the disease	humans cause the disease	humans cause the disease	humans cause the disease
humans have a negative impact on the environment	humans have a negative impact on the ecosystem	humans have a negative impact on the environment	humans have a negative impact on the environment
humans require water to survive	humans require water to survive	humans require water for survival	humans require water for survival
humans produce offspring	humans produce offspring	humans eat plants	humans produce offspring
humans have lived on earth	humans live in the solar system	humans live in the solar system	humans live in the biosphere
humans use fossil fuels for energy	humans use fossil fuels to make energy	humans use fossil fuels to make energy	humans use natural gas to make energy
humans eat green plants	humans eat green plants	humans eat green plants	humans eat green plants
humans eat fruit	humans eat fruit	humans eat fruit	humans eat fruit
humans sometimes eat plants or animals	humans sometimes eat plants and animals	living things sometimes eat insects / animals	animals sometimes eat seeds from trees
a plant absorbs light energy for photosynthesis	a plant absorbs sunlight for photosynthesis	an flower requires energy to grow and provide warmth to the skin	a plant absorbs light for photosynthesis
a plant absorbs water from the air into its roots	a plant absorbs water from the air into its body	a leaf absorbs water from the air through the leaves	a plant absorbs water and nutrients from the air
a plant uses energy to grow	a plant requires energy for growth	a plant requires energy to grow	a plant requires energy to grow
plant reproduction occurs in the spring	plant reproduction occurs in the spring	plant reproduction begins during seed dispersal	plant reproduction begins in spring
plants require water and sunlight to grow	plants require water and sunlight to grow	plants require sunlight to grow and survive	plants require water and sunlight to grow
a plant requires a habitat for survival	a plant needs a habitat for survival	a plant requires a habitat for survival	a plant requires a habitat for survival

Table 24: Explanation reconstruction. From left to right are augmented explanations, decoded explanations from AutoEncoder, explanations from unsupervised INN, and that from supervised INN, respectively.