Learning Disentangled Semantic Spaces of Explanations via Invertible Neural Networks

Anonymous ACL submission

Abstract

Most previous work on controlled text generation have concentrated on the style transfer task: modifying sentences with regard to markers of sentiment, formality, affirmation/negation. Disentanglement of generative factors over Variational Autoencoder (VAE) spaces has been a key mechanism for delivering this type of style transfer control. In this work, we focus on a more general form of controlled text generation, targeting the modification and control of more general semantic features. To achieve this, we introduce a flow-based invertible neural network (INN) mechanism plugged into the Optimus-based AutoEncoder architecture to deliver better properties of separability. Experimental results demonstrate that the model can conform the distributed latent space into a better semantically disentangled space, resulting in a more general form of language interpretability and control when compared to the recent state-of-the-art language VAE models (i.e., Optimus).

1 Introduction

Most previous work on controlled text generation have concentrated on the style transfer task: 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 control (John et al., 2019; Bao et al., 2019; Vasilakes et al., 2022). However, it has been mainly contained in disentangling task-specific(coarse-grained) linguistic factors, especially in style transfer tasks.

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,

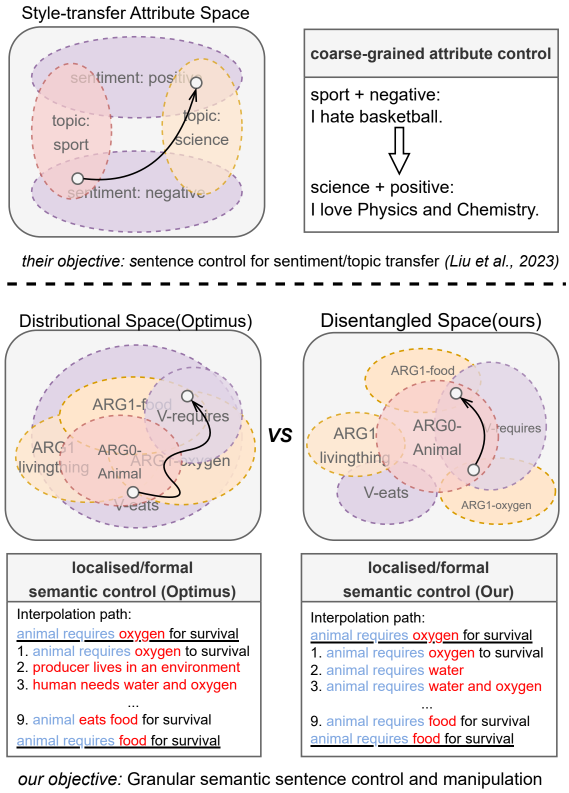

Figure 1: Top: attribute space geometry. Bottom: general semantic geometry, where left: distributional semantic space of Optimus (Li et al., 2020b), right: our compositionality-induced semantic space where the sentence vectors can be located by the intersection of role-content clusters.

where a BERT (Devlin et al., 2018) encoder and a GPT2 (Radford et al., 2019) decoder are connected within a VAE setting. Using representations of conceptually dense explanatory sentences (Jansen et al., 2018b), they showed that sentences, such as *animal requires oxygen for survival*[1], can be represented within a space which can be organised around the associations between predicate, arguments and their associated token content:

[1]Inflections are absent from the dataset's sentences.

ARG0-animal or *VERB-requires*, is geometrically resolved to a hypersolid over the latent space. Nevertheless, the ability to learn and control such separation is still limited as different token-level semantics are still overlapped and entangled in the latent space (e.g., *V-eats* and *V-requires* in Figure 1 bottom left), indicating distributional sentence semantics cannot be currently localised and controlled from the perspective of formal semantics (i.e., *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 predicate-argument structures and their associated content. This type of representation can provide the foundation to shorten the gap between deep latent semantics and formal/symbolic representations (Gildea and Jurafsky, 2000; Banarescu et al., 2013), bridging the flexibility of distributional-neural models with the properties of linguistic grounded representations (e.g. frame/symbolic representations), facilitating both inference interpretability and safety controls.

To deliver this semantic/symbolic control via the distributional sentence space, following the methodological framework introduced by (Zhang et al., 2022), we focus on improving the semantic separability of sentences by focusing on explanatory sentences [2], rather than synthetic or style transfer datasets (Hupkes et al., 2020; Yanaka et al., 2021), in which *compositionality* can be ensured and isolated. 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 AutoEncoder (BERT-GPT2) and the smooth Gaussian space of the INN bottleneck (Figure 3). Specifically, we first pre-train an AutoEncoder to learn sentence representations. Then, we freeze the AutoEncoder and train the INN with sentence representations. Since INN models a bijective transformation, we can control the offline AutoEncoder generation by manipulating the INN latent spaces, which is more efficient and has lower computational demand than re-training a large VAE.

[2]The rationale for choosing explanatory sentences is that they are designed for formal/localised/symbolic semantic inference task in natural language form (Zhang et al., 2023a), which provides a semantically complex and yet controlled experimental setting, containing a both well-scoped and diverse set of target concepts, sentence structures, providing a semantically challenging yet sufficiently well-scoped scenario to evaluate the syntactic and semantic organisation of the space.

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 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. Since the approach leads to a more separable and geometrically consistent sentence space, it can be later operated over to improve the control of the generation of the autoencoder using geometric operators, such as traversal (Higgins et al., 2017) and interpolation (Bowman et al., 2016) (Section 4.2). The contributions of this work are summarised below:

1. We frame the sentence semantic disentanglement from a definition of *compositionality* for bridging formal semantics and distributional representations. **2.** We find that integrating a flow-based INN mechanism into the Optimus architecture is an effective mechanism for transforming the hidden space of the autoencoder into a smooth multivariate Gaussian latent space for representing sentences. **3.** We propose a supervised training strategy for INNs to learn a controllable semantic space with higher disentanglement than previous work. **4.** We use this representation to support semantically coherent data augmentation (controllably generating sentences with well-defined semantic and syntactic properties).

2 Preliminaries

In this section, we first define sentence semantics disentanglement and then illustrate the flow-based INN mechanism and the rationale for its selection.

Sentence semantic disentanglement In view of the *principle of compositionality* (Frege's principle), sentence semantics can be seen as consisting of word-level semantics, which can be jointly represented by word content and its corresponding syntactic/semantic role. In the context of this work, we simplify and particularise this relationship as (*role-content* pair), where the structural syntactic/semantic relationship is defined by its shallow semantics, i.e. as the composition of the content of tokens and their semantic role labels (SRLs). Therefore, this work uses the notion of sentence semantic disentanglement as the cluster separation of the content under SRLs, rather than the notion of feature-dimension binding, common in image disentanglement (Bengio, 2013).

Formally, a sentence s consists of a sequence of different semantic roles (predicate-argument structures and associated types) and word content associations. After encoding in latent space, the semantics of each sentence representation can be described from *general linguistic compositionality*:

$$sem(s) = \underbrace{w_1(c_1, r_1)}_{i.e., ARG0-animal} \oplus \cdots \oplus \underbrace{w_i(c_i, r_i)}_{PRP-survival}$$

where $w_i(c_i, r_i)$ represents the semantics of word i with content c_i (i.e., *animal*) and SRL r_i (i.e., *ARG0*) in context s (i.e., *animal requires oxygen for survival*), $\oplus$ represents *compose* operation. If the sentence representation can be semantically disentangled, the $sem(s)$ can be decomposed into:

$$\begin{aligned} sem(s) &= \{w_i(c_i, r_i)\} \\ &\quad + \{w_1(c_1, r_1) \oplus \cdots \oplus w_1(c_{i-1}, r_{i-1})\} \\ &= \{w_i(c_i, r_i)\} \oplus \{w_1(c_1, r_1)\} \\ &\quad \oplus \{w_2(c_2, r_2) \oplus \cdots + w_1(c_{i-1}, r_{i-1})\} \end{aligned}$$

where each set represents a specific role-content cluster (as illustrated in Figure 2), in this case, given a set of N sentences with the same $w(c, r)$ (i.e., *V-requires*) but different $sem(s)$, those sentence vectors can represent $w(c, r)$ features independently of other features (i.e., *ARG0-animal*), forming $w(c, r)$ cluster. That is, this set of sentence semantics can be composed as:

$$\{sem(s_1), \ldots, sem(s_N)\} = \{w(c, r)\}_{\times N} \oplus \{\ldots\}$$

Therefore, we can evaluate the disentanglement (separability) of sentence semantics by evaluating the density within $\{w(c, r)\}$ set(cluster) (classifier recall) and the separation between different $\{w(c, r)\}$ set(clusters) (classifier accuracy) (as illustrated in section 4.1). Next, we will introduce the INN-based mechanism to learn this semantically disentangled space.

Invertible Neural Networks Flow-based INNs (Dinh et al., 2014, 2016) are a class of neural networks that model the bijective mapping between the observation distribution $p(x)$ and 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

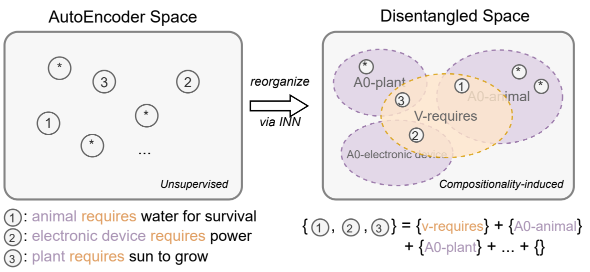

Figure 2: In semantically disentangled space, sentence vectors, ⊛, can be located by the intersection of role-content clusters.

distribution to multivariate Gaussian distributions, INNs use multivariate Gaussian exactly. They can be trained by the following objective function:

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

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 easily performed via the *coupling* layer (Dinh et al., 2014; Kingma and Dhariwal, 2018).

The rationale for choosing flow-based INN is that since it learns the bijective transformation between latent and observed spaces, we can plug-and-play the offline autoencoder generation by manipulating the INN latent space, which is more efficient and has lower computational demand than re-training a large language VAE. Besides, flow-based INNs that learn the prior distribution (i.e., Gaussian) exactly can avoid the information loss from variational inference (ELBO in VAE) where the prior is approximated from posterior $P(z|x)$.

3 Proposed Approach

We encode each sentence x with a frozen autoencoder (i.e., Bert-GPT2) and consider its sentence representation $E(x)$ as the input of INNs (Figure 3). Next, we propose two training strategies to map the hidden representations into Gaussian space.

3.1 Training Strategy

Unsupervised INNs Firstly, we train the INN-based model in an unsupervised fashion, which minimises the negative log-likelihood of the

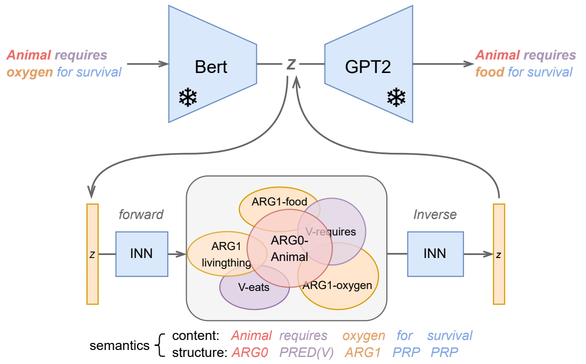

Figure 3: Transforming the representations of explanatory sentences from AutoEncoder, specifically using the same setup as Optimus (Bert-GPT2), into compositionality-induced semantically separable latent space of INN, where a sentence representation can be decomposed into token-level semantics (role-content).

marginal distribution of latent representation $z = E(x)$:

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

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 of geometric clustering property of its latent space by traversal and interpolation (Li et al., 2020b).

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 Optimus autoencoder setting. Using the same foundation, we next 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 Optimus model. For example, given a sentence "***an** animal requires food for survival*" and its central vector of *ARG1-animal*, the training moves the sentence representation closer to the *ARG1-animal* region centre in the INN latent space. Specifically, during the calculation of the posterior, we replace the mean and variance of 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 $z \sim N(\mu_{cluster}, 1-\sigma^2)$. The σ^2 is a parameter which can be empirically determined (in this particular context the optimal value was found be 0.6). More details are provided in Appendix A.

3.2 Data Augmentation

To better capture the different features between distinct role-content clusters, more training sentences are needed in those clusters. Therefore, we consider vector arithmetic and traversal as a systematic mechanism to support data augmentation, which is described in Equations 1.

$$\begin{aligned} vec &= average(E(s_i), E(s_j)) \\ vec[i] &= N(0,1) \quad \forall i \in \{0,.., size(vec)\} \\ s &= D(vec) \end{aligned} \tag{1}$$

where $s_k \in S$ (corpus), $E(s) : S \to \mathbb{R}^n$ is the encoder (embedding) function, and $D(e) : \mathbb{R}^n \to S$ is the decoder function. The term $vec[i] = N(0,1)$ is introduced to resample each dimension and $s = D(vec)$ generates a new sentence. Table 1 lists some randomly selected examples from augmented explanations. Full details on the augmentation algorithm are provided in Appendix A.

Role-content	Augmented sentences
ARG0-animal	an animal requires energy to move some adult animals lay eggs an animal requires shelter an animal can use its body to breathe
ARG0-human	humans travel sometimes humans usually use gasoline humans use coal to make food humans depend on pollinators for survival
PRED-are	wheels are a part of a car toxic chemicals are poisonous green plants are a source of food for animals copper and zinc are two metals
PRED-mean	summit mean the top of the mountain colder mean a decrease in heat energy cleaner mean (less ; lower) in pollutants friction mean the product of a physical change

Table 1: Example of augmented explanations.

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/symbolic 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 of our supervision approach on different semantic roles, including *ARG0*, *ARG1*, *PRED(V)*, where each category has four different word contents, and the same content (i.e., *animal*) with different semantic roles, including *ARG0,1,2*. Reconstructed examples for both unsupervised and cluster-supervised INNs are provided in Appendix D.

Disentanglement between *ARG0* clusters For *ARG0*, we choose *human*, *animal*, *plant*, and *something* according to their 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 aspects (visualisation and classification metrics). Figure 4 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 on the center of their cluster, and there is a clear boundary between clusters, indicating better disentanglement than the baseline models (Optimus, unsupervised INNs).

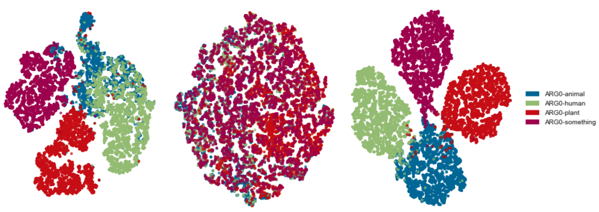

Figure 4: ARG0: t-SNE plot, different colour represents different content regions (blue: animal, green: human, red: plant, purple: something) (left: Optimus, middle: unsupervised, right: cluster supervised). Supervised embeddings concentrate on the respective cluster center.

It is also observable that there are low-density embedding regions at the transition (connection) between two clusters. We decode the middle datapoints between *animal* and *human* clusters and list them in Table 2. 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.

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 2: Middle explanations between *ARG0-animal* and *ARG0-human*.

Next, we 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. Our proxy disentanglement experiments measured the capacity of the classifier to fit the datapoints, thus assessing model separability in-distribution (minimal separability). Therefore, they were evaluated only on the training data. As shown in table 3, all classifiers trained over supervised latent representations outperformed unsupervised INN and Optimus, indicating that the cluster-supervised approach leads to better disentanglement.

ARG0: disentanglement proxy metrics					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.983	0.983	0.983	0.983
	U	0.972	0.972	0.972	0.972
	C	**0.986**	**0.986**	**0.986**	**0.986**
NB	O	0.936	0.936	0.936	0.936
	U	0.961	0.961	0.961	0.961
	C	**0.979**	**0.979**	**0.979**	**0.979**
SVM	O	0.979	0.979	0.979	0.979
	U	0.975	0.975	0.975	0.975
	C	**0.981**	**0.981**	**0.981**	**0.981**

Table 3: 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 4. We can observe that both unsupervised and supervised cases can achieve high invertibility ratios, indicating that the INN mechanism provides the means to control the sentence decoding step precisely by operating the vector over its transformed latent space.

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

Disentanglement between *ARG1* clusters Next, we consider four ARG1 clusters, including *ARG1-food*, *ARG1-oxygen*, *ARG1-sun*, *ARG1-water*, and evaluate model performance following the same procedure. Figure 5 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

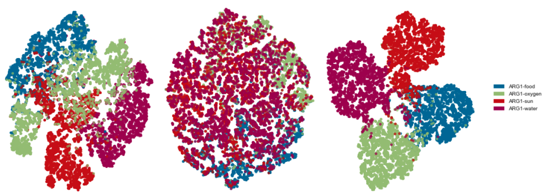

Figure 5: ARG1: t-SNE plot (blue: *food*, green: *oxygen*, red: *sun*, purple: *water*) (left: Optimus, middle: unsupervised INN, right: cluster-supervised INN). Supervision induces separability comparable with ARG0.

5 and 12 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).

Disentanglement between *PRED* clusters Moreover, we analyze the disentanglement between *predicate(PRED)* clusters. Figure 6 shows the distribution of four *PRED* clusters, including *is*, *are*, *cause*, and *require*, over latent space. 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.

ARG1: disentanglement proxy metrics (forward: T)					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.958	0.958	0.958	0.958
	U	0.951	0.951	0.951	0.951
	C	**0.969**	**0.969**	**0.969**	**0.969**
NB	O	0.907	0.907	0.907	0.907
	U	0.926	0.926	0.926	0.926
	C	**0.956**	**0.956**	**0.956**	**0.956**
SVM	O	0.956	0.956	0.956	0.956
	U	0.953	0.953	0.953	0.953
	C	**0.958**	**0.958**	**0.958**	**0.958**

Table 5: Forward evaluation for ARG1, consistent results on different classifiers indicate that supervision can perform better semantic disentanglement.

In Table 6, the supervised INN model achieves better disentanglement and both unsupervised and supervised could obtain a higher ratio. We also evaluate the results for *ARG1* clusters. The same observation holds for both *ARG0* and *PRED*, with details provided in Appendix B.

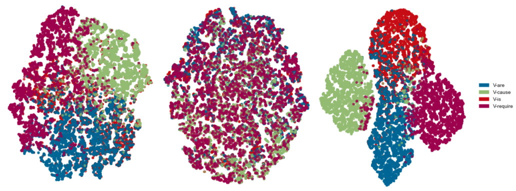

Figure 6: PRED: t-SNE plot (blue: are, green: cause, red: is, purple: require) (left: Optimus, middle: unsupervised, right: cluster supervised).

PRED: disentanglement proxy metrics (forward: T)					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.964	0.964	0.964	0.964
	U	0.959	0.959	0.959	0.959
	C	**0.972**	**0.972**	**0.972**	**0.972**
NB	O	0.923	0.923	0.923	0.923
	U	0.927	0.927	0.927	0.927
	C	**0.951**	**0.951**	**0.951**	**0.951**
SVM	O	0.956	0.956	0.956	0.956
	U	0.950	0.950	0.950	0.950
	C	**0.958**	**0.958**	**0.958**	**0.958**

Table 6: Forward evaluation for predicate clusters, the invertibility ratio is provided in Table 13.

Disentanglement between *ARG0,1,2* clusters The experiments up to this point investigated the separation between the same semantic roles but different content clusters. Next, we explore separating different semantic roles with the same content. We thus focus on the *animal* cluster, and investigate the disentanglement between *ARG0-animal*, *ARG1-*

animal, and *ARG2-animal*. As illustrated in Figure 7, the animal clusters with different semantic roles can be separated after cluster-supervised training, which indicates that the INN model can capture the difference between the same content with different semantic roles in the case of similar topic. That is to say, the INN-based approach could jointly learn separable embeddings w.r.t. role-content and content alone. Table 7 and 14 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 unsupervised and supervised cases can achieve an invertibility ratio of at least 90%.

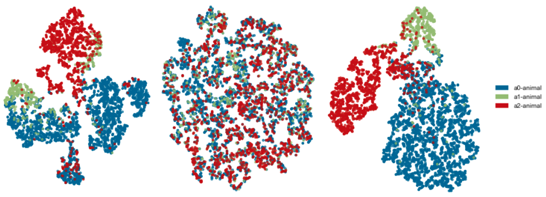

Figure 7: Animal: t-SNE plot (blue: ARG0-animal, green: ARG1-animal, red: ARG2-animal) (left: Optimus, middle: unsupervised, right: cluster-supervised).

Animal: disentanglement metrics (forward: T)					
classifier	train	accuracy	precision	recall	f1 score
KNN	O	0.968	0.968	0.968	0.968
	U	0.960	0.960	0.960	0.960
	C	**0.968**	**0.968**	**0.968**	**0.968**
NB	O	0.929	0.929	0.929	0.929
	U	0.915	0.915	0.915	0.915
	C	**0.940**	**0.940**	**0.940**	**0.940**
SVM	O	0.951	0.951	0.951	0.951
	U	0.931	0.931	0.931	0.931
	C	**0.952**	**0.952**	**0.952**	**0.952**

Table 7: Forward evaluation for Animal, the invertibility ratio is reported in Table 14. Results indicate consistent separation improvement across role clusters.

4.2 Disentanglement Decoding Evaluation

Finally, we evaluate the localised/symbolic generation control of our approach via latent interpolation. It 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 this experiment, we chose the unsupervised INN and Optimus as baselines[3].

In terms of a qualitative characterisation, Table 8 provides the interpolation path of unsupervised INN, cluster-supervised INN, and Optimus, as for the unsupervised INN, we can observe that the intermediate explanations could transition smoothly from source to target for argument. E.g., moving from *humans* to *nonhumans* to *marine animals* to *animals*. However, the *predicate* is changed redundantly, indicating less *predicate-content* disentanglement (i.e., *predicate-require* and *predicate-eat*). Instead, 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 17 and 18.

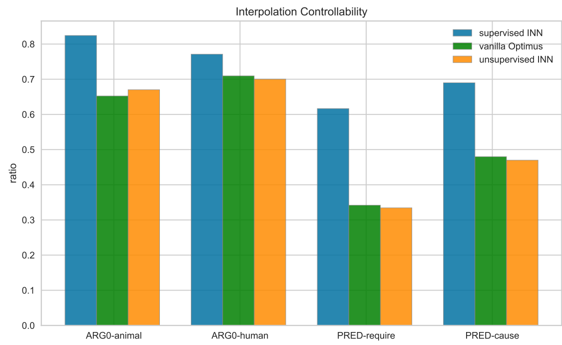

Figure 8: Interpolation control evaluation, we can observe that supervised INN with better semantic separability can lead to better localised semantic control.

Next, we quantitatively evaluate the localised controllability 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. In Figure 8, we can observe that the intermediate sentences from supervised INN can better hold the same role-content as inputs, especially for *predicate(verb)* which usually has a lower effect on distributional sentence semantics (Zhang et al., 2022), indicating our supervision can lead to better latent space separability and localised/symbolic semantic control.

[3]the standard transformer-based VAE(Optimus) with single sentence representation (i.e., the prior is standard Gaussian distribution). Some variant large VAEs, such as Della (Hu et al., 2022), DPrior (Fang et al., 2022), (Li et al., 2022), etc., were not included due to differing training objectives. Additionally, Li et al. (2020b) have illustrated that Optimus can induce smoother interpolations than the Bert-GPT2 autoencoder. Therefore, we don't compare it in our work.

interpolation control: *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
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
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 8: Interpolation examples, indicating the cluster-supervised INN can provide better localised/symbolic semantic control. We also report the interpolations of AutoEncoder in Table 16.

5 Related Work

Sentence Disentanglement In the domain of natural language generation, most previous investigations explored the disentanglement of natural language between two specific linguistic perspectives, such as sentiment-content (John et al., 2019), semantic-syntax (Bao et al., 2019; Zhang et al., 2023b), and negation-uncertainty (Vasilakes et al., 2022), or syntactic-level disentanglement (Mercatali and Freitas, 2021; Felhi et al., 2022), In this work, we focus on general sentence semantics disentanglement from *compositionality* with the target of formal semantic control. This work is the first integration of flow-based INN to support sentence semantics disentanglement.

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) focused on sentence-level representation learning, transforming sentences from a BERT sentence space to standard Gaussian space, which improves sentence embeddings on a variety of semantic textual similarity tasks. Ding and Gimpel (2021) deployed flow-based INN to enrich VAE prior distribution. Gu et al. (2023) use flow to control attributes in style transfer task. This work proposes a supervised training strategy to improve semantic separability, geometrical operations and control over the distributed representation of sentences. Moreover, this work is the first to explore this mechanism to support semantically coherent data augmentation.

6 Conclusions and Future Work

This work focused on the localised/symbolic semantic control of latent sentence spaces, aiming to bridge formal and distributional semantics. We define the sentence semantic disentanglement from the perspective of *compositionality* mapping to the invertibility and bijection properties of INNs. Experimental results indicate that the invertibility mechanisms can transform the distributed hidden space of an autoencoder into one where 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 utilize these geometric properties and semantic controls to support a semantically coherent and controlled data augmentation.

Since our work connects distributional and formal semantics via semantic disentanglement, one possible direction is to apply the same mechanism to explore the safety and control of the formal semantic properties of Large Language Models(LLMs). Besides, recent work (Liu et al., 2023) revealed that disentangled factors can be composed by modelling the moving of latent vectors via ordinary differential equations, which can be adapted in explanatory sentences to explore semantic inference control (i.e., polarity in natural logic (Angeli and Manning, 2014)).

7 Limitations

This work explores how flow-based INN autoencoders can support better formal semantic separation for sentence representations over continuous sentence spaces from the perspective of *compositionality*. While this work is motivated by providing more localised distributed representations, which can positively impact the safety and coherence of generative models, **1.** the specific safety guarantees of these models are not fully established, which we will focus on next. **2.** Additionally, the efficient traversal (sampling) of latent sentence spaces to exert control over generation remains a challenge, particularly given the discrete properties of sentence spaces. **3.** Moreover, the unsupervised INN exhibits a distinct learning pattern for semantic distribution, a topic that requires further explanation in future research. **4.** Furthermore, this study exclusively focused on explanatory sentences, as detailed in (Dalvi et al., 2021), effectively capturing formal semantics for multi-hop natural language inference. However, the exploration of its performance on other types of natural languages is yet to be undertaken.

References

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 534–545, Doha, Qatar. Association for Computational Linguistics.

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.

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.

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*.

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.

Xianghong Fang, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Dit-Yan Yeung. 2022. Controlled text generation using dictionary prior in variational autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 97–111, Dublin, Ireland. 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*.

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.

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. 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.

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.

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*.

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.

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

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.

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.

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

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.

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

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

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.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

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.

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, Ian Pratt-Hartmann, and Andre Freitas. 2023a. Towards controllable natural language inference through lexical inference types. *arXiv preprint arXiv:2308.03581*.

Yingji Zhang, Marco Valentino, Danilo S Carvalho, Ian Pratt-Hartmann, and André Freitas. 2023b. Graph-induced syntactic-semantic spaces in transformer-based variational autoencoders. *arXiv preprint arXiv:2311.08579*.

A Experiment setting

Datasets Table 9 displays the statistical information of the datasets used in the experiment. The data of the two data sets partially overlap, so only the non-repetitive explanations are selected as the experimental data.

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

Table 9: Statistics from explanations datasets.

Table 10 illustrates the semantic, structure, and topic information of explanatory sentences over the latent space (Zhang et al., 2022). Compared with other datasets, such as Wikipedia and Wordnet, that focus on word knowledge, it is more limited, leading to better semantic and structure separability. Table 11 the annotated semantic role categories and corresponding statistic information.

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 neighbours over the latent space.

Algorithm 1 Data Augmentation

Define: R as the role-content set (e.g., ARG1-animal).
Define: S as the explanation corpus (sentences).
Define: V as mapping $\{R \to (S, S)\}$.
Define: $E(s) : S \to \mathbb{R}^n$ as encoder (embedding) function.
Define: $D(e) : \mathbb{R}^n \to S$ as the explanation decoded from Decoder D.
for all $(s_i, s_j) \in V$ **do**
  $vec = average(E(s_i), E(s_j))$
  for all $vec[i] \in vec$ **do**
    $vec[i] = N(0, 1)$ # *resample each dimension*
    $s = D(vec)$ # *new sentence*
  end for
end for

Autoencoder In this work, we employ an autoencoder architecture with the same configuration as described in (Li et al., 2020b)[4]. 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 regularization 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:

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

Figure 9 provides a visual representation of the connection between BERT and GPT2 within the AutoEncoder architecture.

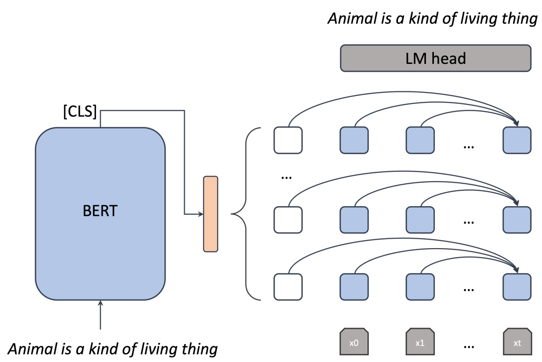

Figure 9: Latent sentence injection.

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 10 displays one single invertible block. The model was implemented using the FrEIA library (Ardizzone et al., 2018-2022) [5]. 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

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

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

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 10: Semantic, structure, topic information of explanatory sentences, where the cluster is the categories of k-means classifier.

the affine coupling block, the sub-network is MLP with 512 as the hidden dimension. Thirdly, we use AdamW (Loshchilov and Hutter, 2017) to optimize the model where the learning rate is 5e-04 in the experiment.

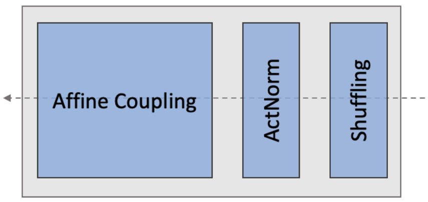

Figure 10: INN one single block.

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

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

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

$$\begin{aligned} y_a, y_b &= \text{split}(y) \\ \log s, t &= m_\theta(y_b) \\ s &= \exp(\log s) \\ x_a &= (y_a - t)/s \\ x_b &= y_b \\ y &= \text{concat}(x_a, x_b) \end{aligned} \tag{3}$$

B Additional supervised INN results

Table 12, 13, and 14 report the invertibility test for *argument1*, *predicate*, and *Animal* clusters, respectively.

Table 15 shows the decoded explanations traversed around the central point of each cluster in the latent space of cluster-supervised INN.

C Controlled Interpolation

In table 17 and 18, 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.

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 11: Semantic Role Labels that appear in explanations corpus. The annotation is done via pretrained model (Shi and Lin, 2019), which can be implemented via AllenNLP library (Gardner et al., 2018).

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

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 13: 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 14: 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 15: Traversal in each cluster (top: *ARG0-Animal*, middle: *ARG1-Animal*, bottom: *ARG2-Animal*).

D INNs: Explanation Reconstruction

Table 19 shows some generated explanations from AutoEncoder and unsupervised INN. As we can see, they can reconstruct the explanations with good quality.

Table 20 shows some reconstructed explanations

Interpolation control: *predicate-require*

source: humans require freshwater for survival

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

target: animals require food to survive

Table 16: AutoEncoder: interpolation examples where top and bottom sentences are source and target, respectively.

Interpolation control: *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 17: Interpolation examples (top: supervised INN, bottom: Optimus).

from AutoEncoder, unsupervised INN, and supervised INN, respectively.

Interpolation control: *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 18: Interpolation examples (top: supervised INN, bottom: Optimus).

Explanations	BERT-GPT2	unsupervised INN
a fish is a kind of organism	a fish is a kind of organism	a fish is a kind of organism
a galaxy is a kind of celestial body	a galaxy is a kind of celestial body	a galaxy is a kind of celestial body
water is the solvent	water is the solute	water is the solvent
metal fork is made of metal for eating	metal fork is made of metal and usually made of metal	metal fork is made of metal for cooking
to carry something means to contain something	to carry something means to bring something	to carry something means to transport that something
a tape measure is a kind of tool for (measuring distance ; measuring length)	a tape measure is a kind of tool for measuring (length ; distance)	a scale is a kind of tool for measuring weight / length
riding something is a kind of movement	walking is a kind of moving	riding is a kind of movement
if a living thing is destroyed then the resources used by that living thing will become available	if something is dead then that something can rest in the environment	if a living thing is destroyed then the resources it uses will be available
The chemical symbol for argon is Ar	The chemical symbol for argon is Ar	The chemical symbol for argon is Ar
exercise has a positive impact on a the strength of a body	strength has a positive impact on a human's survival	strength has a positive impact on a person's health
laying eggs is a kind of property of an animal	laying an egg is a kind of inherited characteristic in birds	laying eggs is a kind of adaptation for reproducing
bears eat berries ; insects ; animals	bears eat berries / insects / animals / food	bears eat berries / insects / animals / berries
pollutants have a negative impact on the (environment ; air quality)	pollution has a negative impact on the (environment ; the environment's water quality ; the environment's resources	pollution has a negative impact on the (environment ; human health)
if an object touches something then one is exposed to that something	if an object touches something then one is exposed to that something	if an object touches something then one is exposed to that something
a stopwatch is a kind of tool for measuring time	a stopwatch is a kind of tool for measuring time	a stopwatch is a kind of tool for measuring time

Table 19: Explanation reconstruction (left: original explanations from WorldTree corpus, middle: explanations from AutoEncoder, right: explanations from unsupervised INN).

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