PAED: Zero-Shot Persona Attribute Extraction in Dialogues

Luyao Zhu, Wei Li, Rui Mao, Vlad Pandelea and **Erik Cambria**
Nanyang Technological University, Singapore
{luyao001, wei008}@e.ntu.edu.sg,
{rui.mao, vlad.pandelea, cambria}@ntu.edu.sg

Abstract

Persona attribute extraction is critical for personalized human-computer interaction. Dialogue is an important medium that communicates and delivers persona information. Although there is a public dataset for triplet-based persona attribute extraction from conversations, its automatically generated labels present many issues, including unspecific relations and inconsistent annotations. We fix such issues by leveraging more reliable text-label matching criteria to generate high-quality data for persona attribute extraction. We also propose a contrastive learning- and generation-based model with a novel hard negative sampling strategy for generalized zero-shot persona attribute extraction. We benchmark our model with state-of-the-art baselines on our dataset and a public dataset, showing outstanding accuracy gains. Our sampling strategy also exceeds others by a large margin in persona attribute extraction.

1 Introduction

Persona attribute extraction in dialogues (PAED) is a crucial task for persona-based dialogue systems (Zheng et al., 2020; Cao et al., 2022). It can extract persona attribute information from conversations. Then, a dialogue system can use extracted persona attributes to generate personalized, user-preference-aware responses to user queries. Previous works define the task as a sentence-level (Daniulaityte et al., 2016) or utterance-level (Gu et al., 2021) classification task. A model learns to classify whether a text contains persona information or not. However, the identified persona-informed texts are still unstructured, resulting in the lower utility of the extracted information in downstream dialogue systems, e.g., irrelevant contexts and different representations towards the same persona attribute. Thus, we define persona attribute extraction as a triplet extraction task. A model should extract a subject, an object, and the persona-relevant relation linking the subject and object from utterances.

The extracted attributes should be in the form of triplets (s, r, o), where the relation (r) indicates the persona attribute type of the subject (s) towards the object (o). Although the existing relation triplet extraction (RTE) task aims to extract triplets from documents (Li and Ji, 2014; Chia et al., 2022), its framework cannot be directly transferred to PAED task because the sentences in documents describe the facts or knowledge in the real world and each pair of entities in triplets can be connected by very limited relations. For example, entities 'Eiffel Tower' and 'France' are very likely connected by relation *located_in* in traditional RTE datasets (Chen and Li, 2021). But the subject and object in dialogues can be linked by many relations, causing hard sample problems, e.g., the relation between 'I' and 'my father' may be *live_with*, *raised_by*, *get_along*, etc. Hence it is essential to formulate a framework for PAED, which is capable of processing hard samples.

To the best of our knowledge, Wu et al. (2020) proposed the largest persona attribute triplet extraction dataset in dialogues based on the triplet annotation of Dialogue Natural Language Inference (NLI) dataset (Welleck et al., 2019). However, we observe that the relation labels were not well-defined in the Dialogue NLI dataset and the dataset of Wu et al. (2020) that inherits the same label set, containing unspecific relation types. For example, negative expressions such as `never`, and `don't have`, are collectively categorized as the relation type of *other* in both datasets. Dialogue NLI missed many triplet labels, resulting in inconsistent annotations, while the dataset of Wu et al. (2020) introduced considerably less reliable labels because utterances and triplet labels were automatically paired by greedy rules. Motivated by addressing the unspecific and inconsistent annotation issues of Dialogue NLI and avoiding the unreliable label-utterance pairing of Wu et al. (2020), we aim to deliver a rigorous dataset for PAED.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9771–9787
July 9-14, 2023 ©2023 Association for Computational Linguistics

We source data from Dialogue NLI and PersonaChat (Zhang et al., 2018) datasets, forming a new dataset, termed PersonaExt. We manually correct 1896 triplet labels of the original Dialogue NLI dataset to improve specificity. We use a more conservative strategy to assign triplet labels to utterances to improve label reliability and consistency: Only the triplet selected by both trained classifiers BERT (Kenton and Toutanova, 2019) and term frequency–inverse document frequency (TF-IDF) is assigned to the utterances. We conduct a human evaluation on PersonaExt and the dataset of Wu et al. (2020). It shows improvements of PersonaExt in label specificity and annotation accuracy.

We formulate PAED as a generalized zero-shot learning (GZSL) task because it is common that the training utterances for a model cannot cover all the relation types. The hard sample issue becomes more severe in GZSL setting. Thus, we propose a generation-based framework with a novel hard negative sampling (HNS) strategy for zero-shot PAED. Our HNS strategy consists of a Meta-VAE sampler and a contrastive structured constraint (CSC).

Meta-VAE sampler uses $|R|$ latent variables of variational autoencoder (VAE) (Kingma and Welling, 2014) to represent $|R|$ kinds of utterance distributions under $|R|$ different relations. It pairs an utterance under a certain relation with one under another relation as positive and hard negative samples, if the distance between their distributions is the shortest. CSC is designed to disperse the paired samples in semantic vector space. On average, our framework surpasses the strongest baseline on our PersonaExt by 1.06% and on the public FewRel dataset (Han et al., 2018) by 0.8% (single triplet) & 3.18% (multiple triplets); Our Meta-VAE sampler exceeds others (Eberts and Ulges, 2020; Yuan et al., 2021b; Zeng et al., 2021) in PAED by 2.66%.

The main contributions of this work are: (1) We develop a PAED dataset, PersonaExt, with 1,896 re-annotated triplets and 6,357 corrected utterance-triplet pairs. (2) We present a generation-based framework for zero-shot PAED. A novel HNS strategy, Meta-VAE sampler with CSC, is presented to enhance the performance of our model. (3) Our model achieves better results than strong baselines in zero-shot PAED and negative sampling. Our code and data are publicly available[1].

2 Related Work

2.1 Persona Extraction

Persona extraction was initially formalized as a classification task inferring user attributes such as gender (Ciot et al., 2013), age (Alekseev and Nikolenko, 2016), opinion (Li et al., 2023), occupation (Preoţiuc-Pietro et al., 2015) and preference (Cambria et al., 2022) from social media. Welleck et al. (2020) formulated persona extraction as a natural language inference task by learning the relation between an utterance and a persona description. Recently, Wu et al. (2020) formalized persona extraction as a generation task extracting structured and easy-to-use user attributes from human-agent dialogues through a two-stage extractor. However, the extractor is not designed for a zero-shot setting.

2.2 Relation Triplet Extraction

RTE was defined as jointly extracting relations (He et al., 2023) and entities (Li and Ji, 2014). Many existing models (Gupta et al., 2016; Zhang et al., 2017; Geng et al., 2021) cannot generalize to unseen relations, which is inevitable in PAED. Chia et al. (2022) proposed a framework RelationPrompt for RTE in a zero-shot setting. However, the above models are tailored for documents and cannot be directly used for PAED as there are more hard samples in dialogues, e.g., the subject and the object may have multiple possible relations. We need to handle the hard samples for zero-shot PAED.

2.3 Hard Negative Sampling

Negative sampling has been proven a key ingredient for contrastive learning (Robinson et al., 2020; Du et al., 2021) and deep metric learning (Suh et al., 2019). Many RTE methods (Qin et al., 2018; Yuan et al., 2021b; Eberts and Ulges, 2020; Guo et al., 2022; Chen et al., 2022) also benefit from robust negative sampling strategies. Hard negative samples that are close to the positive ones in feature space have a crucial influence on extraction performance. As opposed to computer vision-related works (Shrivastava et al., 2016; Liao and Shao, 2022), HNS is rarely studied in RTE and zero-shot settings. Existing joint RTE samplers (Eberts and Ulges, 2020; Yuan et al., 2021a; Zeng et al., 2021) were not designed for hard samples. Therefore, we develop an HNS strategy employing VAE to select hard negative samples to improve the representation ability of the extractor.

[1] https://github.com/SenticNet/PAED

3 PersonaExt Construction

Our PersonaExt dataset is developed for PAED, constructed from multi-turn chitchat datasets, e.g., PersonaChat (Zhang et al., 2018) and Dialogue NLI (Welleck et al., 2019). We use PersonaChat as the data source because it is a dialogue corpus with many personal profiles. PersonaChat was built by two crowd-workers to chat with each other, conditioned on predefined personas, e.g., food preference, job status, and education. Each persona includes 4 to 6 persona sentences. The dataset contains 1,155 personas with over 10,907 dialogues.

Dialogue NLI annotated triplet (s, r, o) for dialogue utterances u and persona sentences p in PersonaChat. They generated *entailment*, *neutral*, or *contradiction* labels for pairs (p, u) and (p, p) based on annotated triplets. For instance, *I adopted a cat* and *I have a cat named Alfred* are both labeled with a triplet (I, *have_pet*, cat). Then, they are considered as *entailment*; Sentences with different triplets are regarded as *neutral* or *contradiction*.

However, many utterances in Dialogue NLI do not have triplet labels, although the utterances contain persona information. Wu et al. (2020) employed a greedy method to improve the label coverage of Dialogue NLI. A triplet label of persona sentence p or utterance u_i is assigned to another utterance u_j if an entailment relationship of (p, u_j) or (u_i, u_j) is predicted by either BERT (Kenton and Toutanova, 2019) or TF-IDF classifiers.

The triplets of Dialogue NLI and the dataset of Wu et al. (2020) are somewhat unreliable, having issues in the unspecific label set definition and inconsistent utterance-triplet pairing. Thus, in our PersonaExt construction, an automatic intersection assignment strategy and manual attribute triplet label correction are used to improve label qualities.

3.1 Automatic Intersection Assignment

Instead of accepting all the triplet labels given by the greedy selection method (Wu et al., 2020), we conservatively assign the triplet label of p or u_i to an utterance u_j only if both BERT and TF-IDF indicate that (p, u_j) or (u_i, u_j) are in an entailment relationship. This change largely improves the label reliability, although there are chances our method may miss a few labels. We believe that extracting reliable persona information is more practical in real-world applications because a dialogue system should avoid using misinformation, although some persona information is conservatively ignored.

3.2 Attribute Triplet Label Correction

We re-annotate the relation types and entities of the attribute triplets in Dialogue NLI, because Dialogue NLI has issues in consistency and specificity. **Consistency.** Some details in persona sentences do not appear in utterances, even if they are in an entailment relationship. For instance, persona sentence *I have 1 cat and I dislike dogs* has the triplets (I, *have_pet*, 1 cat) and (I, *dislike*, dogs); Given an utterance *I usually play with my cat* and the persona sentence are predicted as entailment, the utterance is assigned with the two triplets in Dialogue NLI. However, (I, *dislike*, dogs) did not really appear in the utterance. Thus, the utterance is over-annotated in Dialogue NLI.

Specificity. A relation type should be specific to distinguish it from others. Most negations, such as `never`, and `don't have` are categorized into the relation *other* in Dialogue NLI. However, we expect them to be *never_do* and *have_no*, so that we can categorize them into different personas, e.g., the negation of action and the negation of possession. Besides, an object should be quantity-specific, thus dialogue systems can precisely present the nuances in responses. For example, (I, *have_pet*, 1 cat) and (I, *dislike_animal*, dogs) have more details that can be used to generate responses than (I, *have_pet*, cat) and (I, *dislike*, dogs).

Therefore, we design a semi-automatic annotation method for triplet label correction. **Step 1.** We retrieve persona sentences with negations or any relation type in [*other*, *have*, *like*, *like_general*, <*blank*>], and manually re-annotate them. In total, 1,896 sentences are corrected. The detailed rules and all the relation types in our dataset are in Appendix D. **Step 2.** We assign the triplets of persona sentences to each utterance according to the method in § 3.1. **Step 3.** We use SnowballStemmer[2] to eliminate over-annotations, e.g, redundant numbers, groundless adverbs, adjectives or nouns, and incorrect form of verbs, to make the subject and object consistent with the utterance. The number of processed sentences adds up to 6,357.

In Step 1, we first invited an expert (a main annotator) to manually annotate triplet labels for 1,896 persona sentences. The expert is a native English speaker with rich persona-based dialogue system research experience. We invited a single main annotator to annotate data to secure annotation consistency.

[2] `https://www.nltk.org`

	Object		Relation	
	Cns.	Spec.	Cns.	Spec.
Wu et al. (2020)	0.70	0.68	0.68	0.54
PersonaExt	0.97	0.95	0.89	0.83

Table 1: Evaluation on utterance triplet annotation.

Next, we invited two side annotators to vote triplet labels given by the main annotator and Dialogue NLI dataset, respectively. The main and side annotators follow the criteria that the triplet information should be completely presented in a persona sentence or an utterance; the relationship and object of a triplet should be specific for representing the persona information.

Cohen (1960)'s kappa of the side annotators was 0.72. 82.8% annotations, generated by the main annotator were supported by both side annotators. 89.4% of newly generated annotations were supported by at least a side annotator. We use the newly generated triplet labels that were supported by at least a side annotator. We use the triplet labels of Dialogue NLI if no side annotator supports the re-annotated triplets.

To evaluate the quality of the semi-automatically generated triplet labels in our PersonaExt dataset, we invited two English-speaking graduate students to score 150 randomly selected utterances with triplets. Both the dataset of Wu et al. (2020) and our re-annotated triplets were scored in terms of 'consistency' (cns.) and 'specificity' (spec.) for relations and objects, respectively. The scale of the score was $\{0, 1\}$. The average scores in Table 1 show that PersonaExt largely advances the dataset of Wu et al. (2020) in the two evaluation indices.

4 Generalized Zero-Shot PAED

We propose a framework for PAED in generalized zero-shot learning (GZSL) setting (Xian et al., 2018). The framework consists of two parts: a persona attribute generator (PAG) and a persona attribute extractor (PAE). PAG is trained to generate synthetic dialogue utterances containing persona descriptions. PAE is trained on the synthetic data and extracts attribute triplets of unseen target data. PAE is a pretrained language model (PLM) based extractor enhanced by our proposed Meta-VAE sampler with CSC loss.

4.1 Task Definition

A PAED dataset is denoted as $D = (U, Y)$, where U is the input dialogue utterance set and Y is the persona attribute set. $y = (s, r, o) \in Y$ is an attribute triplet, where s, o, and r are a subject, an object, and a relation type. The goal of generalized zero-shot PAED is to train the model on seen data D_s and generalize to the unseen test data D_t. During training, D_s and test relation R_t are available (Verma et al., 2018). At test time, the relation search space of the trained model contains both training and test relations ($R_s \cup R_t$), and is even much larger as PAE is generation-based instead of classification-based. $R_s \cap R_t = \emptyset$. A test utterance can be assigned to either a training r_s or test relation r_t, where $r_s \in R_s$, $r_t \in R_t$ (Xian et al., 2018).

4.2 Persona Attribute Generator

Prompt tuning is proven to improve the generalization of PLMs in zero-shot learning (Lester et al., 2021), as it bridges the gap between the pretraining tasks and the downstream tasks (Mao et al., 2023). Thus, we prompt-tune the PLM to synthesize samples D_{syn} based on relations r_t in the unseen test set D_t, following the research of Verma et al. (2018). First, PAG is trained on training data D_s, then prompt-tuned with r_t to generate synthetic data D_{syn}. In the testing phase, given a prompt "RELATION: r", PAG is trained to generate a structured output in the form of "CONTEXT: u, SUBJECT: s, OBJECT: o". During training, PAG is trained with the causal language modeling objective, next word prediction (Bengio et al., 2000).

$$p(x_i|x_{<i}; tp) = PAG(x_{<i}), \qquad (1)$$

where x_i is the i-th token in input tokens "RELATION: r, CONTEXT: u, SUBJECT: s, OBJECT: o". We maximize the probability of current token x_i conditioned on previous tokens $x_{<i}$: $\mathcal{L}_g = \sum_{i=1}^{n} \log p(x_i|x_{<i}; tp)$. Temperature tp (Hinton et al., 2015) adjusts the diversity of generation.

4.3 Persona Attribute Extractor

Similarly, we first finetune the PLM-based PAE on training data D_s, then tune the extractor on synthetic samples D_{syn} generated by PAG. PAE is trained with the seq-to-seq objective (Lewis et al., 2020). Given the prompt "CONTEXT: u", the extractor learns to predict a structured output as "SUBJECT: s, OBJECT: o, RELATION: r".

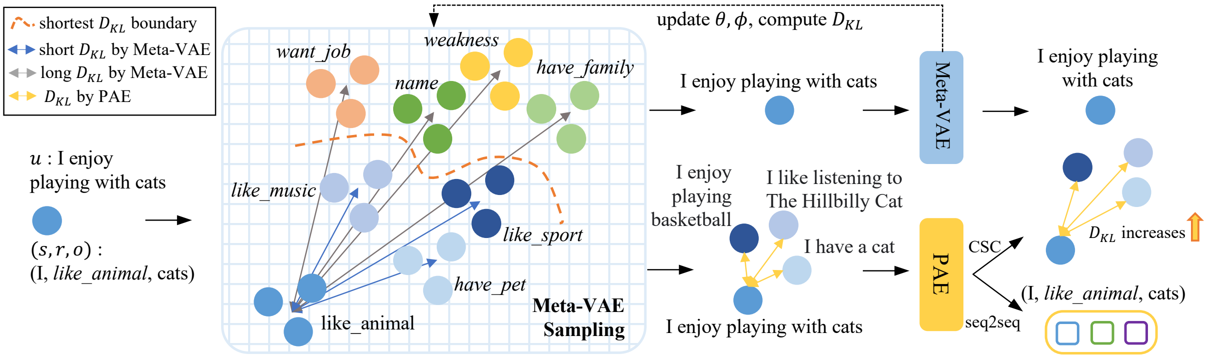

shortest D_{KL} boundary
short D_{KL} by Meta-VAE
long D_{KL} by Meta-VAE
D_{KL} by PAE
u : I enjoy playing with cats
(s, r, o) :
(I, *like_animal*, cats)
want_job
weakness
name
have_family
like_music
like_sport
have_pet
like_animal
Meta-VAE Sampling
update θ, ϕ, compute D_{KL}
I enjoy playing with cats
Meta-VAE
I enjoy playing with cats
I enjoy playing basketball
I like listening to The Hillbilly Cat
I have a cat
I enjoy playing with cats
PAE
CSC
seq2seq
D_{KL} increases
(I, *like_animal*, cats)

Figure 1: Meta-VAE sampler with Contrastive Structured Constraint.

However, during testing, it becomes harder for PAE to distinguish the relation types in unseen data D_t, as a dialogue utterance may convey a completely opposite meaning by replacing only one token, e.g., from ‘like’ to ‘hate’. Hence, we propose CSC to help to differentiate the relation types and Meta-VAE sampler for hard negative sampling, which are introduced in § 4.5 and § 4.4, separately.

4.4 Meta-VAE Sampler

The premise (supported by §4.4.1) of our model is that, for each relation type, Meta-VAE captures the distribution of all the utterances with such a relation. In addition, the utterances u_i with relation r^i and u_j with relation r^j are considered hard negative samples for each other if r^i and r^j is close in terms of distribution distance.

In Fig. 1, an utterance u (*I enjoy playing with cats*) with a triplet (s^+, r^+, o^+) (I, *like_animal*, cats) is a sentence with relation class *like_animal*. And a positive sample of CSC is formulated as “CONTEXT : I enjoy playing with cats . SUBJECT : I OBJECT : cats RELATION : like_animal”. Then, the top-k closest relations (*like_music*, *like_sport*, and *have_pet*) to relation *like_animal* are retrieved by Meta-VAE sampler. For each retrieved relation r', e.g., *like_sport*, an utterance, e.g., *I enjoy playing basketball*, is randomly selected. Then, the selected k utterances are assigned with the same triplet (s^+, r^+, o^+) of u to construct hard negative samples. For example, one of the hard negative samples is “CONTEXT : I enjoy playing basketball . SUBJECT : I OBJECT : cats RELATION : like_animal”. Then the extractor is trained to disperse the positive and negative samples in vector space with CSC and seq2seq loss. Meta-VAE is trained with KL divergence and next word prediction loss as Eq. 4.

4.4.1 Meta-VAE

VAE (Kingma and Welling, 2014) can approximate the prior distribution $p_\theta(z)$ of latent continuous random variable z through approximate posterior $q_\phi(z|u)$ for a given dataset. Intuitively, for each dataset with a certain relation type r, we want to train a VAE to approximate a different prior distribution of its latent continuous random variable. Thus, we will obtain $|R|$ different VAEs in total; $|R|$ is the number of relation classes. However, this is parameter-inefficient. Therefore, we propose Meta-VAE to reduce the complexity: We map each relation class into a relation embedding $Emb_r(r)$ through a fully-connected layer with parameters τ, concatenating each encoded utterance $Emb_u(u)$ with the corresponding relation embedding and feeding the concatenated features into VAE. This is because the concatenation-based conditioning (Sitzmann et al., 2020) equals a special case of hypernetwork (Ha et al., 2017) which is an emerging branch of meta-learning (Chang et al., 2019), and generates the weights of the layers for different tasks (i.e., relation types).

We use GRUs (Chung et al., 2014) as the encoder and decoder of Meta-VAE. Considering the structure of update and reset gates of GRU, we simplify the concatenation by feeding $Emb_r(r)$ as an initial hidden state of a GRU encoder as Eqs. 2 and 3. It is because additive equals concatenation attention (Luong et al., 2015; Bahdanau et al., 2015).

$$a_1^j = \sigma(W_a Emb_u(x_1) + U_a Emb_r(r))^j \qquad (2)$$

$$c_1^j = \sigma(W_c Emb_u(x_1) + U_c Emb_r(r))^j, \qquad (3)$$

where $Emb_u(x_1)$ is the first token embedding in u. a_1^j and c_1^j are the update gate and reset gate at time step 1 for the j-th GRU unit; W_a, U_a, W_c, U_c are learnable parameters.

The empirical objective of Meta-VAE with Gaussian latent variables z is as follows:

$$\mathcal{L}_h(u;\theta,\phi,\tau) = -D_{KL}(q_{\phi,\tau}(z|u)||p_{\theta,\tau}(z)) + \frac{1}{L}\sum_{l}^{L}\text{log}p_{\theta,\tau}(u|z^{(l)}). \tag{4}$$

For each relation r, a set of latent variable z is obtained from the prior distribution $p_{\theta,\tau}(u|z)$ and the data u is generated by the generative distribution $p_{\theta,\tau}(u|z)$ conditioned on z. $z^{(l)} = q_{\phi,\tau}(z|u) \sim \mathcal{N}(\mu_\tau, \sigma_\tau^2 I)$, $p_{\theta,\tau}(z) \sim \mathcal{N}(0, I)$. $q_{\phi,\tau}(z|u)$ is a probabilistic encoder. θ, ϕ and τ are learnable parameters. L is the number of samples.

4.4.2 Sampling Criteria

As the latent variable model can express the distributions of variables in terms of a small amount of latent variables (Bishop, 1998), the latent variable z^r captures the distributions of utterances with different relations r. Thus, we use KL divergence (Kullback and Leibler, 1951) between the distributions of latent variables z^i and z^j to represent the distances between different utterances with different relation classes r^i and r^j.

We assume that the latent variable z of each relation class obeys a multivariate Gaussian Distribution $z \sim \mathcal{N}(z;\mu,\Sigma)$ and all components of z are independent, i.e., $\Sigma_{i,j} = 0, i \neq j$. Then, for latent variables z^i and z^j of relation classes r^i and r^j, the KL divergence is:

$$D_{KL}(P_i||P_j) = E_{P_i}[\text{log}\frac{P_i}{P_j}] = \frac{1}{2}\{\text{log}\frac{|\Sigma_j|}{|\Sigma_i|} - n + tr(\Sigma_j^{-1}\Sigma_i) + (\mu_j - \mu_i)^T\Sigma_j^{-1}(\mu_j - \mu_i)\}, \tag{5}$$

where P_i, P_j are the probabilities of z^i and z^j. As we assume Σ is a diagonal matrix, Eq. 5 can be simplified as:

$$D_{KL}(P_i||P_j) = \frac{1}{2}\{tr(\text{log}\Sigma_j - \text{log}\Sigma_i) - n + tr(\Sigma_i./\Sigma_j) + (\mu_j - \mu_i)^T./\Sigma_j(\mu_j - \mu_i)\}. \tag{6}$$

Here, ./ is an element-wise division operation on Σ_j through which we obtain $1/\sigma_j^k$ for each diagonal element in Σ_j.

Our sampling strategy is: For each relation class r^i, we randomly select one utterance, feeding it to the trained Meta-VAE and obtaining z^i to represent the distribution of utterances under relation r^i; We compute the distance between the distributions of z^i and z^j as the distance between utterances under r^i and r^j, $i \neq j$. Then, the top-k closest relations are selected for each relation r^i; For any utterance, we randomly select one utterance for each top-k closest relations and get k hard negative samples. The detailed sampling algorithm is in Appendix C.

4.5 Contrastive Structured Constraint

The existing generation-based triplet extraction methods seldom focus on the fact that triplets are supposed to be consistent with the input utterance u (Ye et al., 2021). Additionally, the similar token distribution of some dialogue utterances exacerbates the problem. For example, we aim to extract the attribute triplet like (My mom, *have_pet*, 1 cat) instead of (My mom, *like_animal*, 1 cat) for a given input utterance *My mom has a cat named Kitty*. This is because we believe the former explicitly conveys the fact that the cat belongs to my mother, while the latter does not convey the property of ownership.

To this end, we transform the triplet contrastive learning into a binary classification problem: For the utterance u_t with label (s^+, r^+, o^+), we get k hard samples ($u_{t,1}^-, ..., u_{t,k}^-$) from Meta-VAE sampler; We represent the positive sample as "CONTEXT: u_t, SUBJECT: s^+, OBJECT: o^+, RELATION: r^+" and the j-th negative sample as "CONTEXT: $u_{t,j}^-$, SUBJECT: s^+, OBJECT: o^+, RELATION: r^+". We use the hidden state h_i^+ (h_j^-) of the last input token as the i-th positive (j-th negative) sample semantic representation from PAE and feed it to a fully-connected layer to compute classification logits l. Instead of constraining the samples to converge to a fixed positive or negative polarity (Zhu et al., 2020), we employ CSC to relocate the positive and negative samples and make them diverge from each other. The structural contrastive loss based on KL divergence is:

$$\mathcal{L}_c = -D_{KL}(l^+||l^-) - D_{KL}(l^-||l^+) = -\sum_{i=1}^{L}\sum_{j=1}^{k}\frac{1}{k}(l_i^+\text{log}\frac{l_i^+}{l_j^-} + l_j^-\text{log}\frac{l_j^-}{l_i^+}). \tag{7}$$

Here, l_i^+ is the logits for the i-th positive sample and l_j^- is the logits for the j-th negative sample.

5 Experiments

Besides our PersonaExt (PerExt), we experimented on FewRel to explore the capability of our model in multiple triplet extraction and the potential to

	Samples	Entities	Relations	Length
FewRel	56,000	72,964	80	24.95
PersonaExt	35,078	3,295	105	13.44

Table 2: Dataset statistics.

generalize on zero-shot RTE. Another reason is we do not have another triplet-based PAED dataset to test our model. The statistics are listed in Table 2. We evaluated the performance in multiple triplet extraction with a standard metric Micro F_1 (Paolini et al., 2020), precision ($P.$) and recall ($R.$). For single triplet extraction, we used accuracy (Acc.).

5.1 Datasets

FewRel is built through distant supervision where a set of candidate relations and instances are automatically extracted over Wikipedia and Wikidata, and then human annotation is employed to filter low-quality relations (Han et al., 2018). We follow the same operation as Chia et al. (2022) to make FewRel suitable for zero-shot RTE.

For the two datasets, we randomly select a fixed number of seen and unseen labels during training. The number of unseen label size n is set to three incremental setups $\{5, 10, 15\}$. To obtain consolidated experimental results, we use five different random seeds to repeatedly select five combinations of the seen and unseen labels, yielding five different data folds. Each data fold consists of training, validation and test sets. The test set contains sentences with unseen labels. The validation set contains five labels which are used to select sentences for hyper-parameter tuning. The remaining sentences comprise the training set. With this setting, we ensure training, validation and test sentences come from disjoint label sets.

5.2 Baselines

TableSequence (TS) (Wang and Lu, 2020) is primarily designed for joint learning of named entity recognition and relation extraction.

RelationPrompt (RP) (Chia et al., 2022) is the first to solve zero-shot RTE by prompting PLMs to synthesize relation samples given relation labels.

SpERT (Eberts and Ulges, 2020) transfers the strong negative sampler by concatenating the current utterance of which the triplet is (s^+, r^+, o^+) and any other utterance of which the triplet is (s^-, r^-, o^-). The negative triplet is (s^+, r^*, o^-) or (s^-, r^*, o^+), where r^* is a random relation type.

Unseen	Model	FewRel				PerExt
		Multi			Single	Single
		P.	R.	F1.	Acc.	Acc.
n=5	TS	15.23	1.91	3.40	11.82	-
	RP	20.80	24.32	22.34	22.27	38.95
	OURS	**25.79**	**34.54**	**29.47**	**24.46**	**40.01**
n=10	TS	**28.93**	3.60	6.37	12.54	-
	RP	21.59	**28.68**	24.61	**23.18**	26.29
	OURS	23.31	27.42	**25.15**	22.89	**28.09**
n=15	TS	19.03	1.99	3.48	11.65	-
	RP	17.73	23.20	20.08	18.97	27.25
	OURS	**20.68**	**23.39**	**21.95**	**19.47**	**27.57**

Table 3: The experimental results of triplet extraction.

RSAN (Yuan et al., 2021b) randomly selects several relations different from that of the current sentence.

GenTaxo (Zeng et al., 2021) randomly selects a triplet (s^-, r^-, o^-), then the negative triplet is generated as (s^+, r^+, o^-) or (s^-, r^+, o^+).

5.3 Setups

We used the PLM GPT-2 (Radford et al., 2019) with 124M parameters as PAG and BART (Lewis et al., 2020) with 140M parameters as PAE. Meta-VAE sampler has 2.4M parameters. We first finetuned the models on the training set for 5 epochs and selected the best model parameters based on the validation loss with AdamW (Loshchilov and Hutter, 2018) optimizer. We set batch size as 128 for PAG and 32 for PAE, learning rates as 3e-5 for PAG, 6e-5 for PAE and 0.005 for Meta-VAE, and warm up ratio as 0.2. For each relation, 250 sentences were synthesized by PAG utilizing the relation labels of validation and test set as prompts. Then, we finetuned the PAE again on the synthetic sentences. We employed greedy decoding strategy for single triplet extraction and triplet search decoding (TSD) (Chia et al., 2022) strategy for multi-triplet extraction. More implementation details are in Appendix B.

5.4 Experimental Results

We reported the main results for generalized zero-shot RTE and PAED in Table 3. For each $n \in \{5, 10, 15\}$, we run 5 different data folds 3 times and obtained the average with a significance level of 0.05. We found that OURS surpasses RP (1.06% on average) in all settings on PersonaExt. On FewRel dataset, OURS performs better than RP in most settings.

	n=5	n=10	n=15
OURS	39.91	32.47	23.10
w/o HNS	$38.21_{\downarrow 1.70}$	$31.86_{\downarrow 0.61}$	$22.09_{\downarrow 1.01}$
SpERT	$24.38_{\downarrow 15.53}$	$31.79_{\downarrow 0.68}$	$22.47_{\downarrow 0.63}$
RSAN	$37.41_{\downarrow 2.50}$	$30.65_{\downarrow 1.82}$	$21.67_{\downarrow 1.43}$
GenTaxo	$38.66_{\downarrow 1.25}$	$30.55_{\downarrow 1.92}$	$22.15_{\downarrow 0.95}$
Rand	$37.59_{\downarrow 2.32}$	$30.69_{\downarrow 1.78}$	$22.01_{\downarrow 1.09}$

Table 4: Ablation study. Rand means randomly selecting negative sentences with different relation types. HNS refers to Meta-VAE sampler & CSC.

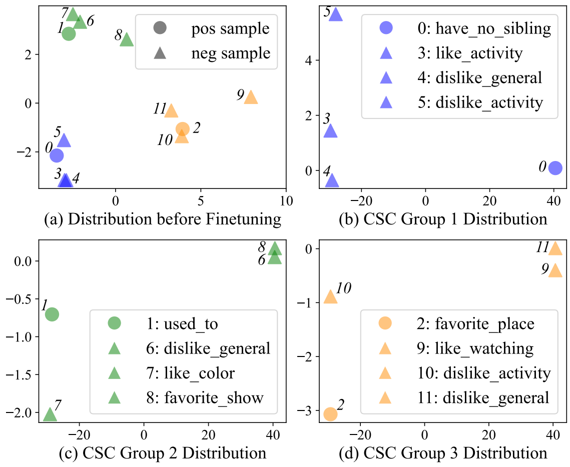
We attribute the significant improvement (3.18% on average) in multi-triplet extraction to our Meta-VAE sampler with CSC that introduces hard samples during training. In particular, OURS consistently achieves higher precision (3.22% on average) than RP. The false positive problem is more severe than the false negative in PAED as speakers are more likely to tolerate negligence rather than confusion. The results also show the generalization capability of our framework on zero-shot RTE.

5.5 Ablation Study

We conducted an ablation study on PersonaExt dataset to compare Meta-VAE sampler with several benchmark samplers. All the samplers use the same random seed and CSC loss. We run them with three unseen label setups and reported the average accuracy of three runs. In Table 4, Meta-VAE sampler outperforms the other four samplers by 2.66% on average and surpasses the strongest baseline GenTaxo by 1.37%. This indicates our Meta-VAE sampler yields better negative samples because of its good approximation to the distributions of different relations. We also observed a significant performance drop of w/o HNS, yet it still exceeds some of the baseline samplers. It suggests a bad sampler may cause a decline instead of an enhancement. Therefore, it is crucial for a sampler to accurately identify the hard negative samples to make the best of contrastive learning.

5.6 Revisiting Meta-VAE Sampler with CSC

The KL divergence between positive and negative samples gets larger during finetuning on the synthetic dataset (details are in Appendix A). This is explained by the fact that we utilized KL divergence to formulate our CSC loss. However, to get a concrete understanding of whether our Meta-VAE sampler and CSC work as expected in vector space, we studied the distribution of positive and negative samples before and after finetuning (Fig. 2).



(a) Distribution before Finetuning

(b) CSC Group 1 Distribution

(c) CSC Group 2 Distribution

(d) CSC Group 3 Distribution

Figure 2: Distribution of positive and negative samples after PCA decomposition. (b), (c), and (d) depict the sample distribution of 3 contrastive groups after finetuning with our Meta-VAE sampler and CSC.

In each contrastive group, one sample is paired with three samples under different relation types retrieved by Meta-VAE. All the scatter points are obtained by decomposing the sample representation h from PAE with principal component analysis (PCA) (Wold et al., 1987). Different groups are in distinct colors. Fig. 2 (a) shows negative samples in each group are closely scattered around the positive sample. This indicates Meta-VAE sampler can find out the hard negative samples which are semantically closest to the positive one. Figs. 2 (b), (c) and (d) suggest fintuning with CSC loss disperses the positive and negative samples in semantic vector space. We conclude that Meta-VAE is capable of retrieving the hard negative samples in terms of semantic meaning and CSC loss enables the model to relocate the positive and negative samples in a sparse manner.

5.7 Case Study

We show three PAED cases in Fig. 3 to reveal the pros and cons of the extraction methods and annotations. As shown, in cases 1 and 3, the RP-extracted objects do not fit well with the relations. In addition, RP extracted incorrect relations which contain the opposite meanings to the ground truth (PersonaExt) in cases 1 and 2. In contrast, the strong performance of our extractor indicates it benefits from dealing with hard negative samples. The object 'all' in case 2 is not specific. Wu et al. (2020)'s annotations of relations and objects in cases 1 and 3 are inconsistent with utterances.

I hate the beach because it will ruin my blond hair.

RP: I, like_general, blond	OURS: I, *dislike_general*, beach
Wu et al.: I, employed_by_general, arena	**PersonaExt**: I, *dislike_general*, beach

I love all sorts of cars, so I just travel the us living in cars.

RP: I, dislike_general, cars	OURS : I, *like_general*, cars
Wu et al.: I, dislike, all	**PersonaExt**: I, *like_general*, cars

I find myself drawn to exotic music which is why I travel to jamaica.

RP: I, like_general, exotic	OURS : I, *favorite_place*, jamaica
Wu et al.: I, like_activity, traveling	**PersonaExt**: I, *favorite_place*, jamaica

Figure 3: Cases of extracted triplets and annotations.

5.8 Exploration of Experimental Settings

To further explore the robustness of our framework, we analyzed the effects of PAE's decoding strategy and the data size of the samples generated from PAG on PersonaExt dataset. The comparison in Table 5 was conducted with three unseen label setups and shows the accuracy change between a decoding strategy and our default greedy strategy. We observed that top-k random sampling (Fan et al., 2018) weakened the extraction performance although it was proved to be more effective than beam search in various generation tasks.

As discussed in Lu et al. (2022), top-k random sampling is commonly used in open-ended generation and, hence, it is not a suitable decoding strategy for PAED. Additionally, TSD improved the accuracy in single triplet extraction in PAED task, which was initially proposed to improve the performance of RP in multi-triplet extraction. However, as TSD is a beam search-based decoding strategy, the slight increase of accuracy came at the cost of much longer computation time. We conducted the experiments on RelationExt dataset with 10 unseen labels and report the results in Fig. 4. In general, the proposed framework is robust with the synthetic data size changing from 250 to 550. An obvious improvement of accuracy can be observed by increasing the synthesized sample number from 1 to 100. The best performance was obtained when the synthesized samples sums up to 450. However, the further increase of the synthetic data size led to gradual reduction in accuracy.

Model	Δ Acc.		
	n=5	n=10	n=15
OURS w/ top-k sampling	-3.66	-2.77	-1.66
OURS w/ TSD	0.54	0.60	0.07

Table 5: Effects of different decoding strategies on single triplet extraction in PAED.

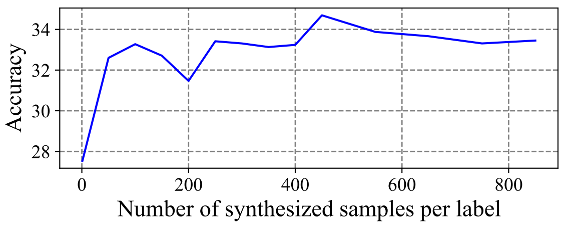


Figure 4: Effects of synthetic data size on PAED.

6 Conclusion

In this work, we studied generalized zero-shot learning for persona attribute extraction in dialogues (PAED). We first built PersonaExt based on PersonaChat and Dialogue NLI through a semi-automatic annotation framework, yielding consistent and specific triplet labels. Then we proposed an effective and interpretable Meta-VAE sampler with CSC loss to process hard negative samples and incorporated it into PAE for generalized zero-shot PAED task. Empirical results demonstrate that our framework surpasses the strongest baseline by a large margin. A visualized quantitative analysis provides a thorough explanation for the mechanism of our Meta-VAE sampler and CSC in sample representations.

Limitations

Due to the lack of theoretical support, it is challenging for us to formalize an annotation scheme for implicit persona attributes in the current stage, e.g., extracting an implicit triplet (I, *like_animal*, dogs) from a sentence "every day, I personally take my dogs out for a walk and lend a hand to my neighbors by occasionally taking their furry friends out for a stroll as well", besides (I, *have_pet*, dogs). Therefore, our PersonaExt is not compatible with the implicit or multiple persona attribute triplet extraction tasks. Additionally, our framework did not exploit complementary information from the context of the current utterance for PAED. For an input with multiple dialogue utterances, it is hard for our model to match extracted persona triplets with the exact speaker because of the existence of pronouns and more than one speaker in dialogues.

Acknowledgements

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

Ethics Statement

In this work, human annotation is conducted with the utmost care to ensure the absence of offensive content and the non-collection of personal identifying information. Comprehensive explanations are provided to the annotators regarding the purpose and appropriate usage of their annotations, ensuring their informed consent has been obtained. The basic demographic and geographic characteristics of the annotator population are not reported, as they do not serve as the primary source of the data in this work.

References

Anton Alekseev and Sergey I Nikolenko. 2016. Predicting the age of social network users from user-generated texts with word embeddings. In *2016 IEEE Artificial Intelligence and Natural Language Conference (AINL)*, pages 1–11. IEEE.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.

Christopher M Bishop. 1998. Latent variable models. In *Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models*, pages 371–403. Springer.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002.

Oscar Chang, Lampros Flokas, and Hod Lipson. 2019. Principled weight initialization for hypernetworks. In *International Conference on Learning Representations*.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. KnowPrompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Raminta Daniulaityte, Lu Chen, Francois R Lamy, Robert G Carlson, Krishnaprasad Thirunarayan, Amit Sheth, et al. 2016. "when 'bad' is 'good'": identifying personal communication and sentiment in drug-related tweets. *JMIR Public Health and Surveillance*, 2(2):e6327.

Bi'an Du, Xiang Gao, Wei Hu, and Xin Li. 2021. Self-contrastive learning with hard negative sampling for self-supervised point cloud learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3133–3142.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Zhiqiang Geng, Yanhui Zhang, and Yongming Han. 2021. Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, 429:132–140.

Jia-Chen Gu, Zhenhua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1126–1136.

Qipeng Guo, Yuqing Yang, Hang Yan, Xipeng Qiu, and Zheng Zhang. 2022. DORE: Document ordered relation extraction based on generative framework. *arXiv e-prints*, pages arXiv–2210.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2537–2547.

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In *International Conference on Learning Representations*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

Kai He, Yucheng Huang, Rui Mao, Tieliang Gong, Chen Li, and Erik Cambria. 2023. Virtual prompt pretraining for prototype-based few-shot relation extraction. *Expert Systems with Applications*, 213:118927.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *stat*, 1050:9.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Max Welling. 2014. Autoencoding variational Bayes. In *International Conference on Learning Representations*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, pages 402–412. Association for Computational Linguistics (ACL).

Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. SKIER: A symbolic knowledge integrated model for conversational emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shengcai Liao and Ling Shao. 2022. Graph sampling based deep metric learning for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7359–7368.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. 2022. NEUROLOGIC A*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2020. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. DSGAN: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769.

Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. 2020. MetaSDF: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147.

Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. 2019. Stochastic class-based hard example mining for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7251–7259.

Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4281–4289.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2020. Dialogue natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3731–3741. Association for Computational Linguistics (ACL).

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Getting to know you: User attribute extraction from dialogues. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 581–589.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Contrastive triple extraction with generative transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14257–14265.

Y Yuan, X Zhou, S Pan, Q Zhu, Z Song, and L Guo. 2021a. A relation-specific attention network for joint entity and relation extraction. In *International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence.

Yue Yuan, Xiaofei Zhou, Shirui Pan, Qiannan Zhu, Zeliang Song, and Li Guo. 2021b. A relation-specific attention network for joint entity and relation extraction. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4054–4060.

Qingkai Zeng, Jinfeng Lin, Wenhao Yu, Jane Cleland-Huang, and Meng Jiang. 2021. Enhancing taxonomy completion with concept generation via fusing relational representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2104–2113.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. End-to-end neural relation extraction with global optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1730–1740.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

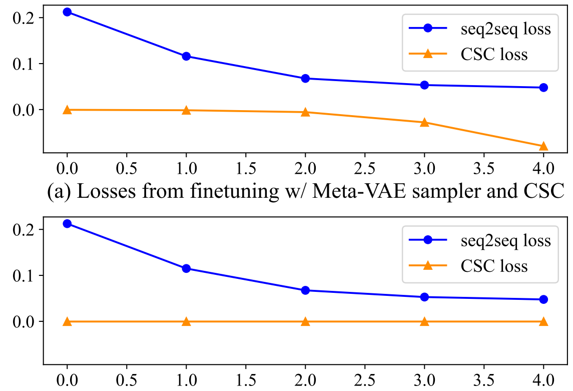
Luyao Zhu, Wei Li, Yong Shi, and Kun Guo. 2020. SentiVec: learning sentiment-context vector via kernel optimization function for sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2561–2572.

A Performance of finetuning with Meta-VAE sampler and CSC

To show the effect of Meta-VAE sampler and CSC during the finetuning process, we report the trend of losses in Fig. 5. Under the same training conditions, we finetuned the persona attribute extractor with or without Meta-VAE sampler and CSC on the synthetic dataset. The losses from the two experiments are depicted in Fig. 5 (a) and (b), separately. The results show that the KL divergence between positive and negative samples selected by Meta-VAE sampler became larger when finetuning with the CSC loss.

B Implementation details

We used one Tesla V100 32 GB GPU for training in our experiments. It took around three minutes to finetune on training set of PersonaExt in each epoch. And it took around two hours for one run of PersonaExt in each setup for each random seed. It took around 5 hours for each run on FewRel dataset. Hyperparameters, i.e., the weight of CSC loss and the learning rate of Meta-VAE are tuned manually according to the performance on the PersonaExt validation set with 5 unseen labels. For the weight of CSC loss, we considered the values 0.5, 0.1, 0.05, 0.01, 0.005; For the learning rate of Meta-VAE, we tried the values 0.05, 0.01, 0.005. Finally, the number of negative samples k is set to 3 and the weight of CSC loss is 0.5. Due to the computational constraints, the other hyperparameters are fixed values, which are listed in Table 6.



(a) Losses from finetuning w/ Meta-VAE sampler and CSC

(b) Losses from finetuning w/o Meta-VAE sampler nor CSC

Figure 5: Loss trends of finetuning on the synthetic dataset.

		Value
Meta-VAE	Dimension of Hidden State	100
	Dimension of Latent Variable	128
	Encoder Layers	2
	Decoder Layers	2
	Bidirectional	True
PAG	Maximum Sequence Length	128
	Sampling Temperature	1.0
PAE	Maximum sequence Length	128
Training	Dropout Rate	0.1

Table 6: Detailed hyperparameters.

Algorithm 1: Meta-VAE Sampler

```
Input: Utterance dataset D, vae-based
       distance function d; number of
       relations $n$; number of negative
       relation sample per relation class $k$
Output: Iterative sampler of dataset D
Initialization: $rel2utt$: dictionary
  containing all utterance indices of each
  relation class;
$utt2rel$: inverse dictionary of $rel2utt$;
$dist$: zero matrix of size $n \times n$.
for $i = 1; i \leq n$ do
    $index_i$=random.choice($rel2utt[i]$)
    $utt_i$ = D[$index_i$]
    for $j = 1; j \leq n$ do
        $index_j$=random.choice($rel2utt[j]$)
        $utt_j$ = D[$index_j$]
        $dist[i, j]$=d($utt_i$, $utt_j$)
    end
end
$dist[i, i]$ = $Inf$
topks = topk($-dist$, $k$)
indices = []
for $idx$ in range($|D|$) do
    sub_index = []
    $relation\_index$=$utt2rel[idx]$
    sub_index.append($idx$)
    for $i$ in topks[$relation\_index$] do
        $select\_utt_i$=random.choice($rel2utt[i]$)
        sub_index.append($select\_utt_i$)
    end
    indices.append(sub_index)
end
return iterator(indices)
```

C Meta-VAE Sampling

The pseudo-code of our Meta-VAE sampler is shown in Algorithm 1.

D Dataset Annotation

D.1 Statistics of Annotated Sentences

In this subsection, we display the statistical information of corrected triplet labels. As shown in Table 7, we manually correct three kinds of errors, i.e., like, neg.(negation), misc.(miscellaneous). Column names 'Like' corresponds to sentences with relations *like* and *like_general*. 'Neg.' corresponds to sentences with negations or zeros in object o. 'Misc.' corresponds to sentences with relations *other*, <*blank*> and *have*. Within the scope of *Automatic*, 'No.' and 'Prn.' refer to the numbers and the pronouns automatically processed by SnowballStemmer, respectively.

	Manual			Automatic	
Type	Like	Neg.	Misc.	No.	Prn.
Count	235	1259	402	1447	4910

Table 7: Statistics of annotated sentences.

D.2 Relation Types

We have 105 relation types in PersonaExt Dataset: live_in_citystatecountry, like_food, place_origin, employed_by_general, like_goto, has_profession, has_age, have_pet, has_ability, never_do, like_music, like_animal, want_do, favorite_food, has_hobby, favorite, like_read, favorite_music_artist, own, employed_by_company, allergy_to, have_vehicle, attend_school, like_drink, favorite_music, have, misc_attribute, previous_profession, dislike_food, physical_attribute, like_sports, school_status, live_with, other, name, favorite_color, belief, like_movie, scared_of, want, favorite_sport, have_children, favorite_hobby, gender, diet, teach, dislike_animal, live_in_general, favorite_animal, have_family, fall_out, dislike_music, do_not_eat, favorite_movie, have_no, job_status, favorite_season, dislike_drink, favorite_activity, worry_about, member_of, do_not_drink, favorite_drink, marital_status, has_degree, favorite_book, do_not_do, dislike_sport, have_children, weakness, international_exp, industry, doing, have_no_family, like_sport, dislike_subject, relationship, like_character, collect, pre_employed_by_company, nationality, sexual_orientation, race, pre_employed_by_general, raised_by, dislike_job, dislike_color, want_no, work_schedule, like_subject, like_activity, like_watching, health_status, favorite_show, dislike_activity, have_no_sibling, used_to, get_along, like_general, have_sibling, dislike_general, like_color, want_job, favorite_place, have_no_children.

D.3 Annotation Rules for Selected Relation Types

The relation types [*other*, *have*, *like*, *like_general*, <*blank*>] are subdivided into the following different relation types based on the semantic meaning of the persona sentence.

- other/ <blank>: {diet, allergy_to, scared_of, relationship, belief, health_status, job_status, school_status, attend_school, doing, used_to, raised_by, work_schedule, get_along, live_with, worry_about, place_origin, race, industry, name, collect, sexual_orientation, misc_attribute, has_ability, have_children, gender, like_music, like_activity, like_goto, like_drink, have_family, have_no_family, live_in_citystatecountry, previous_profession, pre_employed_by_company, physical_attribute, pre_employed_by_general, other}
- have: {collect, relationship, physical_attribute, live_with, live_in_general, like_activity, has_profession, allergy_to, health_status, have_vehicle, international_exp, member_of, want_do, weakness, have_family, has_hobby, marital_status, employed_by, have}
- like/ like_general: {like_character, like_color, like_activity, like_movie, like_music, like_watching, has_hobby, favorite_season, favorite_music_artist, misc_attribute, get_along, job_status, has_profession, collect, have_family, want_job, marital_status, like_general}
- dislike: {dislike_color, dislike_food, dislike_subject, dislike_job, dislike_sport, dislike_animal, dislike_activity, dislike_drink, dislike_read, dislike_music, dislike_general}

Sentences with negations or zeros in o are reannotated by the following relation types based on the context.

- negations/ zeros:{do_not_drink, never_do, have_no_family, have_no, have_no_children, weakness, want_no, have_no_sibling, fall_out, do_not_eat, do_not_do, dislike_job, dislike_food, job_status, scared_of, allergy_to, dislike_color, dislike_sport, dislike_activity, used_to, previous_profession, pre_employed_by_general, marital_status, have_no_pet, health_status, physical_attribute, misc_attribute, sexual_orientation, dislike_general, worry_about, diet, belief, relationship }

E Discussion of data and code

Our PersonaExt is developed on the basis of PersonaChat (MIT license) and Dialogue NLI (CC-BY 4.0). The pretrained language models we used, i.e., GPT-2 and BART, are under MIT license. The data of our PersonaExt is sufficiently anonymized as all persona data are pre-defined instead of extracted information from personal profiles in the real world.

For our human annotation and human evaluation, we invited 5 English-speaking participants among which one is an expert with dialogue system research experience and the other four are graduate students. The hourly payment is around 80% of their hourly salary or stipend. It took totally 64 hours for each annotator in the human annotation task and 5 hours for each in the human evaluation task. The task is scheduled to be finished in one month.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1

☒ A4. Have you used AI writing assistants when working on this paper?
We did not use any AI writing assistant when working on this paper.

B ☑ Did you use or create scientific artifacts?

sections 3 and 4

☑ B1. Did you cite the creators of artifacts you used?
sections 3, 4 and 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We discuss it in the appendix section E.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
In Appendix section E we claimed Human annotation in our work does not show any offensive content or collect any personal identifying information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5 and Appendix section D

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5 and Appendix section D

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5 and Appendix section B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 and Appendix section B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 3 and Sections 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3 and Appendix section E

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix section E

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix section E

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We reported the annotators' research experience and education background in Appendix section E to meet the requirement of the checklist question D2. But we do not report the basic demographic and geographic characteristics of the annotator population and it is not the source of our data.