Contrastive Zero-Shot Learning for Cross-Domain Slot Filling with Adversarial Attack

Keqing He¹*, Jinchao Zhang², Yuanmeng Yan¹, Weiran Xu¹, Cheng Niu², Jie Zhou²
 ¹Beijing University of Posts and Telecommunications, Beijing, China
 ²Pattern Recognition Center, WeChat AI, Tencent
 {kqin, yanyuanmeng, xuweiran}@bupt.edu.cn
 {dayerzhang, withtomzhou}@tencent.com
 {chengniu}@gmail.com

Abstract

Zero-shot slot filling has widely arisen to cope with data scarcity in target domains. However, previous approaches often ignore constraints between slot value representation and related slot description representation in the latent space and lack enough model robustness. In this paper, we propose a Contrastive Zero-Shot Learning with Adversarial Attack (CZSL-Adv) method for the cross-domain slot filling. The contrastive loss aims to map slot value contextual representations to the corresponding slot description representations. And we introduce an adversarial attack training strategy to improve model robustness. Experimental results show that our model significantly outperforms state-of-the-art baselines under both zero-shot and few-shot settings.

1 Introduction

The slot filling task in the goal-oriented dialog system aims to identify task-related slot types in certain domains for understanding user utterances. Traditional supervised slot filling models (Liu and Lane, 2015; Liu and Lane, 2016; Goo et al., 2018; Haihong et al., 2019; He et al., 2020a; He et al., 2020b) have made great achievements. However, these models require massive amounts of labeled data for a new domain, hindering the rapid development of new tasks. To address the data-intensiveness problem, domain adaptation approaches (Bapna et al., 2017; Lee and Jha, 2019; Shah et al., 2019; Obeidat et al., 2019; Liu et al., 2020b; He et al., 2020c) have been successfully applied. In this paper, we focus on zero-shot cross-domain transfer learning which leverages knowledge learned in the source domains and adapts the models to the target domain without labeled training samples in the target domain.

The main challenge of zero-shot slot filling is to identify unseen slot types without any supervision signals in the target domain. Typically, the previous methods rely on slot descriptions or example values to bootstrap to new slots by capturing the semantic relationship between slot descriptions and input tokens. These methods can be classified into two categories: one-stage and two-stage. (Bapna et al., 2017; Lee and Jha, 2019; Shah et al., 2019) conduct one-stage slot filling individually for each slot type. They first generate word-level representations, then interact with the representation of each slot type description in semantic space. Finally, the predictions are independent for each slot type based on the fused features. The main drawback is the multiple prediction problem where a word may be predicted as multiple slot types. In contrast, (Liu et al., 2020a; Liu et al., 2020b) propose a two-stage slot filling framework. They first predict whether the tokens are slot entities or not by a BIO 3-way classifier, then identify their specific slot types based on slot type descriptions. Although the two-stage framework helps learn the general pattern of slot entities, it can't directly leverage auxiliary description information to facilitate to detect BIO labels as the one-stage framework does. Owing to limited labeled training data in source domains, another common issue of zero-shot slot filling is that existing approaches always suffer from weak generalization capability.

Inspired by the above challenges, in this paper, we propose a Contrastive Zero-Shot Learning with Adversarial Attack (CZSL-Adv) method for the cross-domain slot filling. To leverage auxiliary slot

*The work was done when the first author was an intern at Tencent AI Lab. Weiran Xu is the corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.



Figure 1: The overall architecture of our proposed CZSL-Adv model. Following the two-stage setting, we first predict whether the tokens are slot entities or not by a BIO 3-way classifier, then identify their specific slot types based on slot type descriptions. Fig (b) shows the adversarial attack training strategy in the second stage to improve model robustness and Fig (c) displays contrastive representation learning in the first stage to learn the semantic pattern of corresponding slot entities.

description information for detecting BIO labels, we introduce a contrastive learning loss that maximizes the mutual information between raw input encoded representations and corresponding slot description delexicalized input representations. We aim to map slot value representations to the corresponding slot description representations in the latent space. Therefore, slot descriptions can help learn the semantic pattern of related slot entities. To improve the generalization capability of our model, we also propose an adversarial attack training strategy which adds adversarial noise to the inputs in the direction that significantly increases the model's classification loss. The training strategy further improves the adaptation robustness of our method. Our main contributions are three-fold: (1) We propose a contrastive zero-shot learning method for the cross-domain slot filling. (2) We introduce an adversarial attack training strategy to improve model robustness. (3) Experiments on zero-shot learning and few-shot learning settings show that our proposed CZSL-Adv outperforms the state-of-the-art models with large margins. We also provide a comprehensive ablation study and further experiment analysis.

2 Approach

Fig 1 shows the overall architecture of our proposed CZSL-Adv model. In the first stage, it predicts BIO labels with a contrastive representation loss to help learn the semantic pattern of corresponding slot entities. Then in the second stage, it classifies the slot entities into related types with slot descriptions using an adversarial attack training strategy.

2.1 CZSL Model

For a fair comparison, we adopt the same network architecture BiLSTM (Hochreiter and Schmidhuber, 1997) as previous work (Bapna et al., 2017; Lee and Jha, 2019; Shah et al., 2019; Liu et al., 2020a; Liu et al., 2020b). Given an utterance with n tokens as $\mathbf{w} = [w_1, w_2, \dots, w_n]$ and \mathbf{E} denotes the embedding layer for utterances. We formulate the whole process as follows:

$$[h_1, h_2, \dots, h_n] = \text{BiLSTM}(\mathbf{E}(\mathbf{w})) \tag{1}$$

$$[p_1, p_2, \dots, p_n] = \operatorname{CRF}\left([h_1, h_2, \dots, h_n]\right)$$
(2)

where $[p_1, p_2, ..., p_n]$ are the logits for the 3-way BIO classification. Note that we do not show the CRF layer in Fig 1(a) for simplicity. The 3-way BIO classification loss aims to learn the general pattern of slot entities. However, it ignores related slot description representations. Therefore, we introduce a contrastive learning loss to leverage auxiliary slot description information for detecting BIO labels.

Contrastive learning (CL) has achieved great success in the unsupervised visual representation learning (Tian et al., 2019; He et al., 2019; Misra and van der Maaten, 2019; Chen et al., 2020). The main idea

behind CL is to learn representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. In this paper, given a raw input utterance, we generate a positive sample by replacing all the slot entity tokens with the corresponding slot labels. Similarly, replacing the slot entity tokens with different slot labels from the whole slot set will get a set of negative samples. Here, we choose a fixed replaced probability of p = 0.5 for each slot token individually. We randomly sample a minibatch of N examples and define the contrastive loss on pairs of replaced examples derived from the minibatch as follows:

$$\mathcal{L} = \sum_{k=1}^{N} \sum_{i=1}^{M} \max(0, s + d(u_k, u_k^p) - d(u_k, u_{k,i}^n))$$
(3)

where M is the size of the negative sample set and u_k represents k-th input utterance vector in the batch. u_k^p denotes the positive sample and $u_{k,i}^n$ denotes *i*-th negative sample of k-th input utterance. d is the L2 distance function and s is the margin. Following (Felbo et al., 2017; Liu et al., 2020b), we employ another BiLSTM and an attention layer to generate representations of positive and negative samples. The contrastive loss aims to map slot value contextual representations to the corresponding slot description representations in the latent space. Therefore, slot descriptions can help learn the semantic pattern of related slot entities. Compared to Template Regularization (TR) proposed by Liu et al. (2020b), our CZSL jointly models pairs of positive and negative samples to distinguish semantic representations of different slot types.

2.2 Adversarial Attack Training

In this section, we introduce an adversarial attack training strategy as shown in Fig 1(b) to improve model robustness. Firstly, we can obtain a slot description matrix $M_{desc} \in \mathbb{R}^{n_s \times d_s}$ where n_s is the number of all the slot types and d_s is the dimension of slot description representation. Following (Shah et al., 2019), we sum the embeddings of the slot description tokens as the description representation. Then, we perform average pooling over the hidden states for k-th slot entity tokens to get r_k and calculate the dot product as classification logits $s_k = M_{desc} \cdot r_k$. Finally, we can get the classification cross-entropy loss \mathcal{L}_{slot} .

Due to limited labeled data in source domains, existing approaches are always vulnerable to noisy input utterances. Hence, apart from traditional classification entropy loss \mathcal{L}_{slot} , we apply Fast Gradient Value (FGV) (Miyato et al., 2017; Vedula et al., 2020) to approximate a worst-case perturbation as a noise vector:

$$\tilde{\boldsymbol{v}}_{\text{noise}} = \epsilon \frac{g}{\|g\|}; \text{ where } g = \nabla_e \mathcal{L}_{slot}$$

$$\tag{4}$$

Here, the gradient is the first-order differential of the loss function \mathcal{L}_{slot} w.r.t. e, representing the direction that rapidly increases the loss function. We perform normalization and then use a small ϵ to ensure the approximate is reasonable. Then we add the noise \tilde{v}_{noise} and perform the second forward to get a new loss \mathcal{L}'_{slot} . Finally, we use the adversarial attack loss $\mathcal{L}_{adv} = \mathcal{L}_{slot} + \mathcal{L}'_{slot}$ for the backpropagation. Adversarial noise enables the model to handle extensive noisy input utterances and can be regarded as a data augmentation mechanism. Experiments also show that the adversarial training strategy can effectively improve the performance of our CZSL method.

3 Experiment

3.1 Setup

Dataset. To evaluate our approach, we conduct experiments on Snips (Coucke et al., 2018), a personal voice assistant dataset that contains 7 domains and 39 slots, where some slots are shared across domains while the others are domain-specific. In table 1, we give detailed statistics of Snips dataset. For each domain in Snips, we give number of samples, list of cross-domain shared slots, and list of domain-specific slots. To test our framework, each time, we choose one domain as the target domain and the other six domains as the source domains.

Baselines. In our experiments, we compare our approach with the following zero-shot/few-shot slot filling baselines:

Domain	number of samples	slots					
Domani	train/dev/test (total)	cross-domain shared	domain-specific				
AddToPlaylist	1818/100/124 (2042)	artist, playlist, music_item	playlist_owner, entity_name,				
BookRestaurant	1881/100/92 (2073)	country, state, timeRange,	party_size_number, served_dish,				
		sort, spatial_relation, city	restaurant_type, party_size_description,				
		sort, spatial_relation, erty	facility, restaurant_name, poi, cuisine				
GetWeather	1896/100/104 (2100)	country, state, city,	condition_description, geographic_poi,				
	1890/100/104 (2100)	timeRange, spatial_relation	condition_temperature, spacurrent_location				
PlayMusic	1914/100/86 (2100)	year, sort, artist, playlist, music_item	year, album, genre, service, track				
RateBook	1876/100/80 (2056)	object_type, object_name	object_part_of_series_type, rating_value,				
Kalebook	· · · ·	object_type, object_name	object_select, best_rating, rating_unit				
SearchCreativeWork	1847/100/107 (2054)	object_type, object_name	-				
FindScreeningEvent	1852/100/107 (2059)	timeRange, spatial_relation,	object_location_type, location_name,				
	1052/100/107 (2059)	object_type,	movie_name, movie_type				

Table	1:	Detailed	statistics	of Si	nins	dataset.
rabic	т.	Detuneu	statistics	01 01	mps	uuuset.

- **Concept Tagger (CT)** A method proposed in (Bapna et al., 2017), which utilizes slot descriptions (e.g. "date of departure" for slot *date_of_departure*) to boost the performance on detecting novel slots in the target domain.
- **Robust Zero-shot Tagger (RZT)** A method proposed in (Shah et al., 2019), which utilizes both slot descriptions and slot example values (e.g. "iHeart Radio" for slot *service*) for zero-shot slot filling.
- **Coarse-to-fine Approach (Coach)** A method proposed in (Liu et al., 2020b), which splits the crossdomain slot filling task into two stages: coarse-grained BIO 3-way classification and fine-grained slot type classification, and uses slot descriptions in the second stage to help recognize unseen slots.
- **Coach+TR** A variant of Coach, which further applies template regularization to improve the slot filling performance of similar or the same slot types, and achieves better results.

Implementation Details To conduct experiments under zero-shot settings, we follow the set-ups in (Liu et al., 2020b). First, we combine samples from the rest 6 domains for training. Then we split samples in the target domain into two sets: 500 samples as a validation set and the remain as a test set. For few-shot (50 samples) experiments, we further add 50 samples from the target domain to the training set. We fine-tune all hyperparameters in the validation set and report the F1-score in the test set. We achieve best results when s set to 0.15, ϵ set to 0.1 and M set to 2.

We use character-level and word-level embeddings for each input token and the total embedding dimension is 400. We set the hidden size of BiLSTM to 200 and use a dropout rate to 0.3 for all BiLSTM encoders. We set We use Adam optimizer (Kingma and Ba, 2014) to optimize all parameters with a learning rate of 0.0005. We set the batch size to 32 and use the early stop of patience 5.

3.2 Main Results

Table 2 displays the main results of our CZSL-Adv method compared to the state-of-the-art baselines. Our method outperforms the SOTA models by 3.6% on the average F1-score under zero-shot learning setting, and 1.76% under few-shot learning setting. The improvements demonstrate the effectiveness of our proposed method. Besides, CZSL and Adv respectively achieve superior performance by 1.74% and 1.05% under zero-shot learning setting. The contrastive loss can help learn the semantic pattern of related slot entities by corresponding slot descriptions. And the adversarial attack training strategy also achieves significant improvement. We observe our method gets better improvement under the zero-shot learning setting than few-shot. We hypothesize that our CZSL-Adv method effectively alleviates data scarcity than the previous models under the zero-shot learning setting.

3.3 Ablation Analysis

We compare the effect of CZSL and Adv in Table 2. For zero-shot experiments, both CZSL (39.13%) and Adv (38.44%) achieve better average F1-score than previous state-of-the-art (37.39%), which proves both

Training Setting	Zero-shot					Few-shot on 50 samples								
$\overline{\text{Domain} \downarrow \sim \text{Model} \rightarrow}$	CT	RZT	Coach	Coach+TR	CZSL	Adv	CZSL-Adv	CT	RZT	Coach	+TR	CZSL	Adv	CZSL-Adv
AddToPlaylist	38.82	42.77	45.23	50.90	53.29	52.51	53.89*	68.69	74.89	71.63	74.68	77.71	78.81*	76.18
BookRestaurant	27.54	30.68	33.45	34.01	37.97*	33.24	34.06	54.22	54.49	72.19	74.82	77.35*	75.08	76.28
GetWeather	46.45	50.28	47.93	50.47	48.70	48.07	52.24*	63.23	58.87	81.55	79.64	81.85	80.73	83.28*
PlayMusic	32.86	33.12	28.89	32.01	29.14	34.92*	34.59	54.32	59.20	62.41	66.38	65.59	63.55	68.17 *
RateBook	14.54	16.43	25.67	22.06	29.55	27.05	31.53*	76.45	76.87	86.88	84.62	84.31	86.74	87.22
SearchCreativeWork	39.79	44.45	43.91	46.65	49.32	47.63	50.61*	66.38	67.81	65.38	64.56	66.41	64.59	66.49
FindScreeningEvent	13.83	12.25	25.64	25.63	25.95	25.63	30.05*	70.67	74.58	78.10	83.85	81.14	82.46	83.26
Average F1	30.55	32.85	35.82	37.39	39.13	38.44	40.99*	64.85	66.67	74.02	75.51	76.34	75.99	77.27*

Table 2: Slot F1-scores on SNIPS for different target domains under zero-shot and few-shot learning settings. Scores in each row represent the performance of the leftmost target domain. CZSL denotes our proposed contrastive zero-shot learning model and Adv represents the adversarial attack method in the paper. * indicates the significant improvement over all baselines (p < 0.05).

setting	0 san	nple	50 samples			
setting	unseen	seen	unseen	seen		
СТ	27.10	44.18	62.05	69.64		
RZT	28.28	47.15	63.96	73.10		
Coach	32.89	50.78	74.65	76.95		
Coach+TR	34.09	51.93	76.49	80.16		
CZSL	34.57	52.69	77.15	80.09		
Adv	34.32	53.12	77.18	78.93		
CZSL-Adv	36.35	55.43	78.48	79.36		



Table 3: Average F1-scores for seen and unseen slots across all target domains.



CZSL and Adv contribute to the final improvement. When compared to the full model (40.99%), Adv shows severer performance degradation (-2.55%) than CZSL (-1.86%), indicating that the performance improvement comes more from CZSL.

3.4 Seen Slots vs. Unseen Slots

Table 3 shows the results on seen and unseen slots in target domains. We can observe that our CZSL-Adv consistently outperforms the baselines on the unseen slots under the two settings but gets a relatively small drop on seen slots under the few-shot setting. The results prove that our CZSL-Adv makes an effect on the zero-shot learning scenario without sufficient supervised signals.

3.5 Analysis of Norm of Adversarial Attacks

Fig 2 displays the effect of norm ϵ of adversarial noise. ϵ controls the range of adversarial noise \tilde{v}_{noise} . We can see that for different target domains, $\epsilon = 0.1$ always achieves better performance.

4 Conclusion

In this paper, we propose a Contrastive Zero-Shot Learning with Adversarial Attack (CZSL-Adv) method for cross-domain slot filling. The main contributions are contrastive representation learning and adversarial attack training. The former leverages slot descriptions to help learn the semantic pattern of related slot entities and the latter improves model robustness by augmenting noise inputs. Extensive experiments show the effectiveness of our proposed method, especially for the zero-shot learning setting.

Acknowledgments

We thank all anonymous reviewers for their constructive feedback. This work was partially supported by National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701, DOCOMO Beijing Communications Laboratories Co., Ltd.

References

- Ankur Bapna, Gokhan Tür, Dilek Hakkani-Tür, and Larry Heck. 2017. Towards zero-shot frame semantic parsing for domain scaling. *Interspeech 2017*, Aug.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.
- Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.
- Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*.
- Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.
- E Haihong, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning.
- Keqing He, Weiran Xu, and Yuanmeng Yan. 2020a. Multi-level cross-lingual transfer learning with language shared and specific knowledge for spoken language understanding. *IEEE Access*, 8:29407–29416.
- Keqing He, Yuanmeng Yan, and XU Weiran. 2020b. Learning to tag oov tokens by integrating contextual representation and background knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 619–624.
- Keqing He, Yuanmeng Yan, Hong Xu, Sihong Liu, Zijun Liu, and Weiran Xu. 2020c. Learning label-relational output structure for adaptive sequence labeling. In 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation, 9:1735–1780.

- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Sungjin Lee and Rahul Jha. 2019. Zero-shot adaptive transfer for conversational language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6642–6649, Jul.
- Bing Liu and Ian Lane. 2015. Recurrent neural network structured output prediction for spoken language understanding. In Proc. NIPS Workshop on Machine Learning for Spoken Language Understanding and Interactions.
- Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech 2016*, Sep.
- Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020a. Zero-resource cross-domain named entity recognition.
- Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020b. Coach: A coarse-to-fine approach for crossdomain slot filling.
- Ishan Misra and Laurens van der Maaten. 2019. Self-supervised learning of pretext-invariant representations.
- Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. *arXiv: Machine Learning*.
- Rasha Obeidat, Xiaoli Z. Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Description-based zero-shot finegrained entity typing. In NAACL-HLT.
- Darsh Shah, Raghav Gupta, Amir Fayazi, and Dilek Hakkani-Tur. 2019. Robust zero-shot cross-domain slot filling with example values. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open intent extraction from natural language interactions. *Proceedings of The Web Conference* 2020.