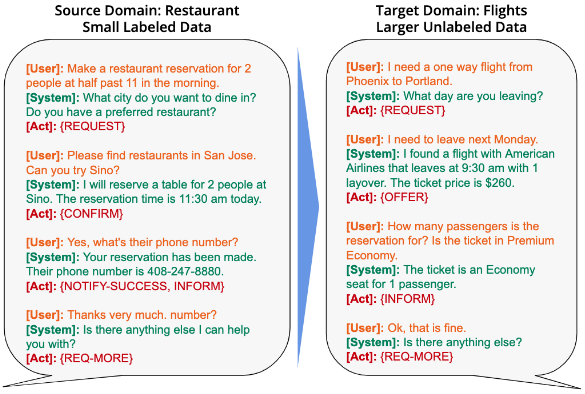
Simple Data Augmentation with the MASK Token Improves Domain Adaptation for Dialog Act Tagging

Semih Yavuz **Kazuma Hashimoto** **Wenhao Liu**
Nitish Shirish Keskar **Richard Socher** **Caiming Xiong**
Salesforce Research
{syavuz, k.hashimoto, wenhao.liu, nkeskar, rsocher, cxiong}@salesforce.com

Abstract

The concept of Dialogue Act (DA) is universal across different task-oriented dialogue domains - the act of "request" carries the same speaker intention whether it is for restaurant reservation or flight booking. However, DA taggers trained on one domain do not generalize well to other domains, which leaves us with the expensive need for a large amount of annotated data in the target domain. In this work, we investigate how to better adapt DA taggers to desired target domains with only unlabeled data. We propose MASKAUGMENT, a controllable mechanism that augments text input by leveraging the pre-trained MASK token from BERT model. Inspired by consistency regularization, we use MASKAUGMENT to introduce an unsupervised teacher-student learning scheme to examine the domain adaptation of DA taggers. Our extensive experiments on the Simulated Dialogue (GSim) and Schema-Guided Dialogue (SGD) datasets show that MASKAUGMENT is useful in improving the cross-domain generalization for DA tagging.



[User]: Make a restaurant reservation for 2 people at half past 11 in the morning.
[System]: What city do you want to dine in? Do you have a preferred restaurant?
[Act]: {REQUEST}

[User]: Please find restaurants in San Jose. Can you try Sino?
[System]: I will reserve a table for 2 people at Sino. The reservation time is 11:30 am today.
[Act]: {CONFIRM}

[User]: Yes, what's their phone number?
[System]: Your reservation has been made. Their phone number is 408-247-8880.
[Act]: {NOTIFY-SUCCESS, INFORM}

[User]: Thanks very much. number?
[System]: Is there anything else I can help you with?
[Act]: {REQ-MORE}

[User]: I need a one way flight from Phoenix to Portland.
[System]: What day are you leaving?
[Act]: {REQUEST}

[User]: I need to leave next Monday.
[System]: I found a flight with American Airlines that leaves at 9:30 am with 1 layover. The ticket price is $260.
[Act]: {OFFER}

[User]: How many passengers is the reservation for? Is the ticket in Premium Economy.
[System]: The ticket is an Economy seat for 1 passenger.
[Act]: {INFORM}

[User]: Ok, that is fine.
[System]: Is there anything else?
[Act]: {REQ-MORE}

Figure 1: Overview of dialog act tagging task and cross-domain generalization scenario of similar dialog acts. The specific contents of the utterances of the same dialog act (DA) are distinct due to the domain difference, making the cross-domain generalization challenging.

1 Introduction

Dialog act (DA) tagging, one of the important NLU components of modern task-oriented dialog systems, aims to capture the speaker's intention behind the utterances at each dialog turn. Several different schema and taxonomies have been introduced by several different researchers (Core and Allen, 1997; Stolcke et al., 2000; Bunt et al., 2010; Mezza et al., 2018) over the years. However, the main focus of the recent work (Kumar et al., 2018; Chen et al., 2018; Raheja and Tetreault, 2019) on DA tagging was on human-human social conversations (Godfrey et al., 1992; Jurafsky et al., 1997), which is less applicable for task-oriented setting.

Recently, several task-oriented dialogue datasets (Shah et al., 2018; Henderson et al., 2014; Budzianowski et al., 2018) have been released. However, the discrepancy in their annotation schema hinders the progress on building DA taggers that can generalize across domains and possibly datasets. To address this issue, Paul et al. (2019) propose a universal schema for DAs by aligning annotations for multiple existing corpora. In this regard, another useful corpora employed as a testbed in this work is Schema-guided dialogues (SGD) (Rastogi et al., 2020), which covers 20 domains under the same DA annotation schema.

It is often challenging and costly to obtain a large amount of in-domain dialogues with annotations. However, unlabeled dialogue corpora in target domain can easily be curated from past conversation logs or collected via crowd-sourcing (Byrne et al., 2019; Budzianowski et al., 2018) at a more reasonable cost. The goal of this work is to investigate how to leverage pre-trained masked language models (e.g., BERT) to better adapt DA taggers to unseen domains with available unlabeled dialogues. Pre-trained language models (Devlin et al., 2019; Liu et al., 2019) have been successful for several NLP tasks including dialogue systems (Wolf et al.,

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5083–5089,
November 16–20, 2020. ©2020 Association for Computational Linguistics

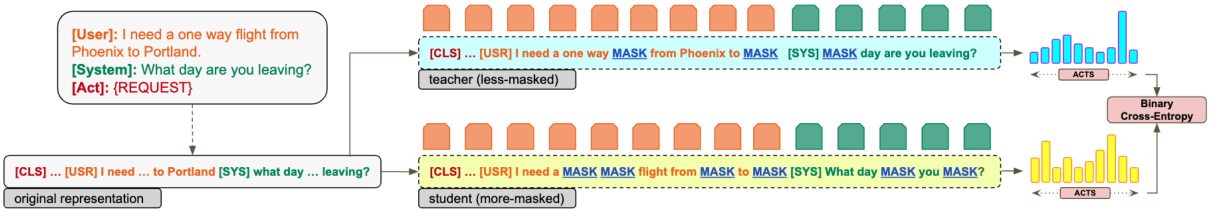

[User]: I need a one way flight from Phoenix to Portland.
[System]: What day are you leaving?
[Act]: {REQUEST}
[CLS] ... [USR] I need ... to Portland [SYS] what day ... leaving?
original representation
[CLS] ... [USR] I need a one way MASK from Phoenix to MASK [SYS] MASK day are you leaving?
teacher (less-masked)
[CLS] ... [USR] I need a MASK MASK flight from MASK to MASK [SYS] What day MASK you MASK?
student (more-masked)
ACTS
ACTS
Binary Cross-Entropy

Figure 2: Given a dialogue turn in target domain, we obtain *teacher* and *student* representations by applying two different maskings on its flattened original representation. We use the output binary probability distributions (per dialog act) of the teacher as soft targets to train the student. Orange and green colored boxes indicate different segment ids.

2019; Zhang et al., 2019; Bao et al., 2020; Henderson et al., 2019; Wu et al., 2020). However, domain adaptation capabilities of these models remain to be further explored for goal-oriented dialogues.

In this paper, we use the pre-trained MASK token of BERT model to define MASKAUGMENT, which stochastically augments text input by randomly replacing its tokens with the MASK token. We adopt consistency regularization approach (Sajjadi et al., 2016) to introduce an unsupervised teacher-student learning scheme by leveraging MASKAUGMENT for generating teacher and student representations retaining different amount of the original content from the unlabeled dialogue example. Our extensive experiments on GSim (Shah et al., 2018) and SGD (Rastogi et al., 2020) datasets suggest: (i) BERT establishes a much stronger baseline compared to previous work (Paul et al., 2019), (ii) The proposed teacher-student learning via MASKAUGMENT is useful in further improving the target domain F1 score over BERT baseline: up to 3% when the full source domain data is used, and up to 10% for the low-resource setting.

2 MASKAUGMENT

In this section, we first discuss the task setup, BERT-based DA tagging model, and relevant background. We then define the proposed fine-tuning objectives leveraging MASKAUGMENT.

2.1 Task Setup

We start by formalizing the DA tagging task, depicted in Figure 1, as a multi-label classification problem. Let $D = [T_1, T_2, \dots, T_n]$ denote a dialogue of n turns as a series of user and system utterances. Let $A = \{a_j\}_1^m$ be the predefined set of m different DAs in the schema. The objective of dialogue act tagging is to determine a subset $A_k \subseteq A$ of DAs that apply to the current turn T_k given the conversation history $D_{:k} = [T_1, T_2, \dots, T_k]$ so far. We formulate this objective simply as a classification problem with binary labels $y_j \in \{0, 1\}$ for each act a_j where $y_j = 1$ if $a_j \in A_k$ and $y_j = 0$ otherwise. As defined above, dialogue act tagging is a turn-level classification problem, hence every turn T_k constitutes: (i) a labeled example $(D_{:k}, A_k)$ if we have a set A_k of DA annotations, or (ii) an unlabeled example $(D_{:k}, \cdot)$ otherwise.

2.2 Model

Given a conversation history $D_{:k}$ as input, we first convert it into a sequence of words by concatenating user and system utterances. Before concatenating each utterance, we prepend it with corresponding speaker tag using [SYS] and [USR] special tokens indicating system and user sides, respectively. Finally, the whole flattened sequence is finalized by prepending it with [CLS] special token to obtain the final *dialogue history representation*:

$$x = [\text{CLS}] ... [\text{USR}]\ T_i\ [\text{SYS}]\ T_{i+1} ... \qquad (1)$$

The segment ids are set to 0 and 1 for the tokens of past turns and the current turn, respectively.

For DA tagging task, dialogue history x is used as input to pre-trained language model M, and the model computes a probability vector $p_\theta(\cdot|x) = \sigma(WM(x) + b)$ where $M(x) \in \mathbb{R}^d$ is the output contextualized embedding corresponding to CLS token, $W \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ are trainable weights of a linear projection layer, σ is the sigmoid function, θ denotes the entire set of trainable parameters of model M along with (W, b), and finally $p_\theta(a_j|x)$ indicates the probability of tag a_j being triggered. The following objective is used to train the model parameters.

Supervised tagging loss (STL). This objective is used to update the DA tagger via the supervision coming from labeled source data S. We use binary-cross entropy loss $\mathcal{J}_{\text{STL}}(\theta; x, y)$ defined as:

$$-\left[y \cdot \log p_\theta(\cdot|x) + (1 - y) \cdot \log(1 - p_\theta(\cdot|x))\right] \quad (2)$$

2.3 Learning with MASKAUGMENT

Semi-supervised learning (SSL) (Berthelot et al., 2019, 2020; Sohn et al., 2020; Li et al., 2020) is an effective approach for improving deep learning models by leveraging in-domain unlabeled data. Unlike traditional SSL setting, our objective is to primarily address the underlying source-to-target domain shift. In prior work (Xie et al., 2019; Wei and Zou, 2019), unsupervised data augmentation methods including word replacement and back-translation have been shown useful for short written text classification. However, such augmentation methods are shown to be less effective (Shleifer, 2019) when used with pre-trained models. Besides, back-translation is less applicable in our scenario as translation of multi-turn dialogue itself is a rather challenging task compared to short text.

Instead, we propose a simple and controllable data augmentation–MASKAUGMENT–to explore a new unsupervised teacher-student learning scheme for domain adaptation of DA taggers. MASKAUGMENT augments the original text input by randomly replacing its tokens with MASK token at a specified probability. We follow the masking policy in (Devlin et al., 2019). Formally, let $z(\bar{x}|x,\epsilon)$ denote the MASKAUGMENT as a stochastic transformation with ϵ-probability for input x. Below we define three fine-tuning objectives leveraging MASKAUGMENT that are used in addition to $\mathcal{J}_{\text{STL}}$.

Masked tagging loss (MTL). We incorporate MASKAUGMENT into the STL objective by perturbing its input sequence x as follows:

$$\mathcal{J}_{\text{MTL}}(\theta; x, y, \epsilon) = \mathbb{E}_{\bar{x}\sim z(\bar{x}|x,\epsilon)}\left[\mathcal{J}_{\text{STL}}(\theta; \bar{x}, y)\right].$$

Masked LM loss (MLM). This is the original objective that BERT is pre-trained with. The objective of MLM training is to correctly reconstruct a randomly selected subset (with probability ϵ) of input tokens leveraging the unmasked context. We denote this loss by $\mathcal{J}_{\text{MLM}}(\theta; x, \epsilon)$.

Teacher-Student Learning with Disagreement Loss (DAL). We adopt consistency regularization (Sajjadi et al., 2016; Laine and Aila, 2017) widely used in traditional SSL (Berthelot et al., 2019; Sohn et al., 2020; Li et al., 2020) and define *disagreement loss*, which employs MASKAUGMENT in a novel way to give rise to an unsupervised teacher-student training. The core idea is to contrast the amount of controllable perturbations to learn more generalizable representations. We propose a stochastic imputation-based teacher and student selection by leveraging MASKAUGMENT. As in Figure 2, we sample two augmentations $\bar{x}^{(t)} \sim z(\bar{x}|x,\epsilon_t)$ and $\bar{x}^{(s)} \sim z(\bar{x}|x,\epsilon_s)$ for teacher and student, respectively. We take $\epsilon_t < \epsilon_s$ to ensure that the teacher augmentation $\bar{x}^{(t)}$ retains more of the original content x than the student augmentation $\bar{x}^{(s)}$, hence is more reliable. The disagreement loss $\mathcal{J}_{\text{DAL}}(\theta; x, \epsilon_t, \epsilon_s)$ is then computed as the binary cross-entropy loss between the teacher $p_\theta(\cdot|\bar{x}^{(t)})$ and the student $p_\theta(\cdot|\bar{x}^{(s)})$ distributions as in Eq. 2, treating teacher as the soft target (y).

3 Experiments

3.1 Datasets

GSIM (Shah et al., 2018) consists of machine-machine task-oriented dialogues in two tasks of two different domains: buying a movie ticket (GMov) and reserving a restaurant table (GRes). It contains 1500/469/1117 dialogues for the train/dev/test sets. Following (Paul et al., 2019), its dialogue acts are mapped to 13 tags in universal schema.

SGD (Rastogi et al., 2020) consists of 22,825 schema-guided single/multi-domain dialogues where domains can have multiple schemas, each defined by a set of tracking slots. We use single-domain dialogues of smaller sizes including music (SMusic), media (SMedia), ride-sharing (SRide) as source domains to study generalization on flights (SFlights), the largest one, as the target domain.

3.2 Training and Implementation Details

The final loss function is the sum of the active ones among $\mathcal{J}_{\text{STL}}, \mathcal{J}_{\text{MTL}}, \mathcal{J}_{\text{DAL}}, \mathcal{J}_{\text{MLM}}$ except $\mathcal{J}_{\text{MLM}}$ is multiplied with 0.1 when active. DAL is activated after 1 epoch of training with the remaining objectives. We perform a tuning of $\epsilon_t \in [0, 0.1]$ and $\epsilon_s \in [0.1, 0.5]$ for DAL objective. We optimize the loss using AdamW (Loshchilov and Hutter, 2017). The learning rate is tuned on $[10^{-5}, 5 \times 10^{-5}]$ with no warmup steps. We use a batch of 16 examples with maximum sequence length of 128, which covers around 9.9, 10.3, 9.9 turns on average for train, dev, test splits, respectively. We use *transformers* library[1] for our implementation.

3.3 Results and Discussion

We begin our discussion with our main findings on domain adaptation as presented in Table 1. We explore the effect of incorporating our proposed MTL and DAL objectives on top of STL (baseline) for both Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2019) models. Transformer baseline model on DA tagging with STL

[1] https://github.com/huggingface/transformers

Fine-tuning Objectives			**GMov → GRes**		**GRes → GMov**		**SMusic → SFlights**		**SMedia → SFlights**		**SRide → SFlights**	
STL	MTL	DAL	Source	Target	Source	Target	Source	Target	Source	Target	Source	Target
LSTM (Paul et al., 2019)			91.4	75.1	89.2	85.0	-	-	-	-	-	-
Transformer												
✓	✗	✗	96.5	81.6	97.4	93.6	84.7	57.7	92.9	76.4	91.5	62.3
✓	✗	✓	96.2	85.4	97.3	93.8	86.0	58.5	91.5	76.6	97.3	62.3
✓	✓	✗	97.0	83.8	96.5	94.1	90.3	58.4	93.0	76.3	96.6	64.6
✓	✓	✓	97.3	85.9	97.6	94.7	92.6	59.8	94.3	78.9	97.4	65.4
scratch-BERT												
✓	✗	✗	98.4	89.7	98.7	96.9	93.6	60.6	98.3	82.9	98.8	67.2
✓	✗	✓	97.8	91.4	99.0	97.1	93.9	60.8	98.0	86.5	98.8	67.5
✓	✓	✗	98.3	90.9	98.9	97.5	95.8	60.8	98.5	84.4	98.5	69.7
✓	✓	✓	98.9	**92.8**	99.0	**97.7**	98.6	**62.6**	98.4	**89.0**	99.3	**71.1**

Table 1: Micro-F1 scores on the test set of source and target domains with combinations of STL, MTL, and DAL objectives. scratch-BERT is initialized from original *bert-base-uncased*. Transformer is a randomly initialized version of scratch-BERT.

Model	**scratch-BERT**	**pre-BERT**
STL	89.7	91.9
STL + MLM	91.0	93.2
STL + MTL + DAL	92.8	94.0
STL + MTL + DAL + MLM	94.1	94.4

Table 2: Micro-F1 scores on target (GRes) domain for pre-BERT (obtained by domain-adaptive pre-training) in comparison with scratch-BERT (initialized from BERT) across different fine-tuning objectives. We also highlight the effect of MLM when used as a fine-tuning objective on unlabeled target domain examples in the second and fourth rows.

	Precision		**Recall**	
Model	Dev	Test	Dev	Test
scratch-BERT				
STL	87.8	88.3	89.6	91.1
STL + MTL + DAL	91.5	90.7	95.3	95.0
pre-BERT				
STL	91.8	91.4	92.1	92.4
STL + MTL + DAL	93.1	92.4	95.6	95.6

Table 3: Precision and recall scores on target (GRes) domain for pre-BERT and scratch-BERT including dev set results.

objective leads to considerable improvements on the LSTM (Paul et al., 2019). Fine-tuning BERT with STL objective from scratch provides further improvements on Transformer, establishing a much stronger baseline both on source and target domain performance. For both Transformer and BERT models, our proposed DAL and MTL objectives are independently useful in further improving the cross-domain generalization over strong baselines that are trained only with STL objective while not hurting the source domain performance. Moreover, fine-tuning on the combined unsupervised objective of DAL and MTL leads to the best performance (last row) on target domains across the board, hinting they provide orthogonal benefits.

Domain-adaptive pre-training (pre-BERT). As shown useful by Gururangan et al. (2020), we explore domain-adaptive pre-training of BERT model on the combination of source and target domain dialogues with MLM loss before fine-tuning it on the task. As presented in Table 2, pre-BERT helps improve the F1 score on the target domain (GRes) by up to 2.2% over the strong scratch-BERT model across different training objectives. Incorporating MASKAUGMENT into pre-BERT via our proposed DAL and MTL objectives leads to 2.1% boost over fine-tuning with only STL, achieving 4.8% F1 score improvement over LSTM (Paul et al., 2019) (89.2%) trained on the full labeled data (GRes) itself in a supervised way. This might partly be due to the effect of learning a more domain-aware MASK token, which in return may lead to a more informed and useful teacher representations.

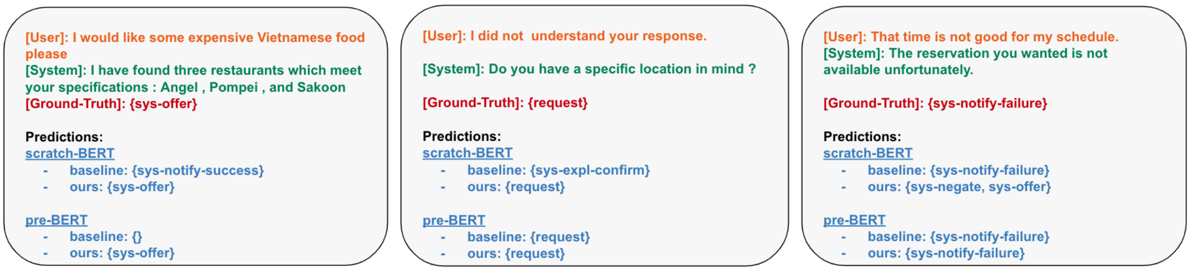
Model	**#Dials: 10**	**#Dials: 20**	**#Dials: 50**
scratch-BERT			
STL	53.3	65.5	73.6
STL + MTL + DAL	58.4	69.0	78.2
pre-BERT			
STL	59.8	73.9	82.9
STL + MTL + DAL	70.4	77.8	85.1

Table 4: F1 scores on target domain (GRes) under the low-resource setting. #Dials denote the number of labeled dialogues (randomly sampled) used in the source domain (GMov). We report the average of 3 runs with different samples.

The effect of MLM in fine-tuning. We also conduct experiments on using MLM as unsupervised fine-tuning objective on the target domain dialogues. As shown in Table 2, it helps improve the cross-domain generalization performance. Specifically, our ultimate model (last row) achieves 94.1% and 94.4% F1 scores on the target domain for scratch-BERT and pre-BERT models, respectively.

Consistent gains on precision and recall. In Table 3, we demonstrate that our proposed approach leads to consistent gains on both precision and recall. While the improvement is consistent, we observe that MASKAUGMENT significantly helps close the recall gap between scratch-BERT and pre-BERT (i.e., from 2.5% to 0.3% on the dev set and from 1.3% to 0.6% on the test set).

Low-resource setting for source domain. As shown in Table 4, we observe that the benefit of



[User]: I would like some expensive Vietnamese food please
[System]: I have found three restaurants which meet your specifications : Angel , Pompei , and Sakoon
[Ground-Truth]: {sys-offer}

Predictions:
scratch-BERT
- baseline: {sys-notify-success}
- ours: {sys-offer}

pre-BERT
- baseline: {}
- ours: {sys-offer}

(a) Improved example for *sys-offer*.

[User]: I did not understand your response.

[System]: Do you have a specific location in mind ?

[Ground-Truth]: {request}

Predictions:
scratch-BERT
- baseline: {sys-expl-confirm}
- ours: {request}

pre-BERT
- baseline: {request}
- ours: {request}

(b) Improved example for *request*.

[User]: That time is not good for my schedule.
[System]: The reservation you wanted is not available unfortunately.

[Ground-Truth]: {sys-notify-failure}

Predictions:
scratch-BERT
- baseline: {sys-notify-failure}
- ours: {sys-negate, sys-offer}

pre-BERT
- baseline: {sys-notify-failure}
- ours: {sys-notify-failure}

(c) Failure example for *sys-notify-failure*.

Figure 3: Qualitative examples comparing baseline and proposed approach across scratch-BERT and pre-BERT settings.

		scratch-BERT		pre-BERT	
Acts	**Frequency**	Baseline	Ours	Baseline	Ours
affirm	13%	92.0	94.3	95.5	94.2
inform	30%	95.3	95.0	95.3	95.8
repeat	1%	95.2	90.9	98.3	89.5
request	15%	92.2	97.8	97.0	99.3
sys-expl-confirm	6%	76.5	87.4	86.8	89.6
sys-negate	3%	89.8	78.2	84.9	82.3
sys-notify-failure	4%	93.8	82.4	85.0	84.5
sys-notify-success	3%	80.7	91.8	95.1	88.2
sys-offer	13%	69.2	89.3	71.5	91.1
thank-you	2%	98.5	85.5	98.5	97.1
user-hi	6%	99.6	99.9	99.7	99.2
user-negate	4%	87.4	88.5	89.8	91.9

Table 5: Micro-F1 scores for each dialog act (DA) on the test split of target (GRes) domain. Note that we use the target data without their labels in totally unsupervised fashion, where only the source (GMov) domain provides label supervision. We compare baseline (STL) and our proposed training scheme (STL + MTL +DAL) through MaskAugment for both scratch-BERT and pre-BERT settings. Frequency indicates the occurrence ratio of the corresponding dialog act in the test split of the target domain. We highlighted the rows with more than 10% frequency. Green highlighting indicates the tags on which our method is superior to baseline, and red highlighting indicates the opposite.

MaskAugment through DAL and MTL objectives becomes larger as the number of labeled dialogues in the source domain gets smaller. The effect of domain-adaptive pre-training also becomes stronger, providing 12% improvement over scratch-BERT when only 10 labeled dialogues are available in the source domain while achieving 85.1% F1 score on the target domain with 50 labeled dialigues when combined with MaskAugment.

Adaptation performance across DAs. In Table 5, we present additional analysis on the adaptation performance across the set of all dialog acts in the schema. MaskAugment provides significant improvement across most of the DAs including frequent ones such as *request* and *sys-offer* while not hurting the performance much (if not improving) on other frequent acts such as *affirm* and *inform*. For scratch-BERT setting, baseline (STL) objective obtains superior performance on less frequent DAs including *sys-negate*, *sys-notify-failure*, and *thank-you*, for which the performance drop is mostly bridged in pre-BERT setting. On the other hand, Pre-BERT provides consistent adaptation improvement over scratch-BERT across all dialog acts except for *sys-negate* and *sys-notify-failure*.

Qualitative analysis of the approach. In Figures 3a and 3b, we provide examples for improved predictions on *sys-offer* and *request* acts, respectively. These are some of the most frequent DAs that MaskAugment can provide a significant (5-20%) improvement over the baseline approach for both scratch-BERT and pre-BERT settings. In Figure 3c, we include an example where scratch-BERT with MaskAugment fails on predicting *sys-notify-failure* act correctly as opposed the baseline. However, most of such failure cases vanish for pre-BERT setting, where the gap in F1 score drops from 11.4% in scatch-BERT to only 0.5% in pre-BERT as shown in Table 5.

4 Conclusion

We study cross-domain generalization of pre-trained language models for DA tagging. While the fine-tuned BERT model performs well on in-domain DA tagging, its cross-domain generalization is still not satisfactory. To combat this shortcoming, we investigate domain adaptation through the proposed unsupervised teacher-student training that leverages the MaskAugment method for data augmentation. Our empirical results show that the proposed training scheme leads to significant improvements on domain adaptation for dialog act taggers. In the future, we plan to explore MaskAugment for other tasks in NLP domain.

Acknowledgments

We thank Xinyi Yang and Tian Xie for their insightful feedback, and the anonymous reviewers for their helpful and thoughtful comments.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2020. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations (ICLR)*.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Harry Bunt, Jan Alexandersson, Jean Carletta, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Mark G Core and James F Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

John J. Godfrey, Edward Holliman, and Jan McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discoursefunction annotation coders manual, draft 13.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Shachi Paul, Rahul Goel, and Dilek Z. Hakkani-Tür. 2019. Towards universal dialogue act tagging for task-oriented dialogues. In *INTERSPEECH*.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Pararth Shah, Dilek Z. Hakkani-Tür, Gökhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak Kennard, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *ArXiv*, abs/1801.04871.

Sam Shleifer. 2019. Low resource text classification with ulmfit and backtranslation. *CoRR*.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *ArXiv*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*.

Ashish Vaswani, Shazeer Noam, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149.

Chien-Sheng Wu, Steven C. H. Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogues. *ArXiv*, abs/2004.06871.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation for consistency training. *arXiv: Learning*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *ArXiv*, abs/1911.00536.