INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Tracking Must Go On : Dialogue State Tracking with Verified Self-Training

Jihyun Lee[1], *Chaebin Lee*[1], *Yunsu Kim*[1,2], *Gary Geunbae Lee*[1,2]

[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
{jihyunlee, leecbin0911, yunsu.kim, gblee}@postech.ac.kr

Abstract

In task-oriented dialogues, dialogue state tracking (DST) is a critical component as it identifies specific information for the user's purpose. However, as annotating DST data requires a significant amount of human effort, leveraging raw dialogue is crucial. To address this, we propose a new self-training (ST) framework with a verification model. Unlike previous ST methods that rely on extensive hyper-parameter searching to filter out inaccurate data, our verification methodology ensures the accuracy and validity of the dataset without using a fixed threshold. Furthermore, to mitigate overfitting, we augment the dataset by generating diverse user utterances. Even when using only 10% of the labeled data, our approach achieves comparable results to a fully labeled MultiWOZ2.0 dataset. The evaluation of scalability also demonstrates enhanced robustness in predicting unseen values.

Index Terms: dialogue state tracking, self-training, augmentation, multi-domain dialogue systems

1. Introduction

The growing interest in artificial intelligence speakers and virtual personal assistants has made task-oriented dialogue (TOD) essential as it aims to achieve the user's purpose through dialogue. The TOD system typically involves multiple models, and among them, the dialogue state tracking (DST) model plays a critical role; it generates a belief state that contains specific information for the user's objective [1]. For instance, in Figure 1, the belief state includes a slot (*hotel-area*) and value (*North*) information that is necessary to meet the user's hotel booking requirements.

In the field of DST research, MultiWOZ [2] has become a widely-used benchmark dataset that includes dialogue data and corresponding gold truth labels. Numerous studies [3, 4, 5, 6, 7] have been conducted based on this dataset. However, creating a realistic DST corpus poses significant challenges as it requires domain knowledge and extensive human labor for manual annotation. Additionally, DST models need to handle newly emerging values, such as hotel names or taxi departures, which need a comprehensive understanding of conversations and avoiding overfitting. This necessitates the inclusion of diverse user utterances with appropriate annotations, which can be a time-consuming process. Consequently, leveraging raw dialogue without annotation, such as audio recognition results or QnA chatting logs, is an essential and practical research area in the field of DST

Our research focuses on addressing the question: 'How can we leverage unannotated data to obtain more extensive information compared to using only labeled data?' To tackle this challenge, we employ a self-training (ST) approach. In ST, a teacher model is trained using a limited amount of labeled data, which is then used to predict pseudo-labels for the unlabeled data. A subset of the pseudo-labeled data is incorporated into the labeled dataset, and the student model is trained using the combined dataset. This self-training method effectively uses the potential of unlabeled data, making it particularly suitable in scenarios where labeled data is scarce, but unlabeled data is abundant [8, 9, 10]. This aligns well with the challenges encountered in DST research.

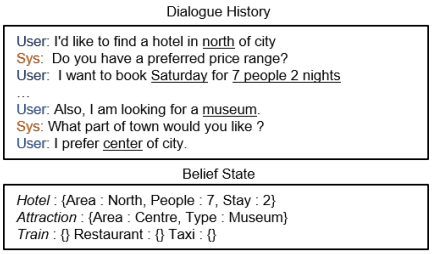

Figure 1: *Example of task oriented dialogue task.*

Previous ST studies face difficulties in selecting an appropriate subset from the pseudo-labeled dataset. Typically, the teacher model's confidence value is used for selecting subset [11, 12, 13, 14], but this method needs to find fixed threshold value, which requires hyper-parmeter searching. To address this, some researchers employ the entire pseudo-labeled dataset [15, 16] which usually performs less than an empirically searched threshold [17, 18, 19]. Furthermore, confidence values are typically calculated on the log-likelihood of the generative model [20, 21, 22], despite of there being no clear relationship between a high log-likelihood and the accuracy of the pseudo-labels [16].

To this end, we propose a novel self-training (ST) framework that incorporates a verification model for DST. The verification model is designed to determine the validity of a dialogue and its pseudo-label pair. By leveraging this model, we could select a more accurate subset from the pseudo-labeled dataset compared to the confidence-based method. Moreover, our proposed method directly utilizes the dialogue and label pair, enabling the generation of a dynamic amount of subset datasets that align with the pseudo label quality of the teacher model. This approach is more rational than using a fixed threshold. Furthermore, to enhance scalability for values that are not present in the training dataset, we have devised an augmentation method. This method generates diverse user utterances by conditioning them on previous history and belief state. To ensure the quality of the augmented dataset, we utilize the trained verification model to select a valid subset from the augmented data.

10.21437/Interspeech.2023-2513

We refer to our proposed ST framework as **LAVe**, which stands for **L**abeling, **A**ugmentation, and **Ve**rification self-training.

In the experiment, LAVe has demonstrated noteworty improvements in the initial performance of the teacher models on the MultiWOZ2.0 dataset, with a rise from 46.58% to 49.75%. Remarkably, even when utilizing only 10% of the labeled data, LAVe achieves results comparable to those of a fully labeled dataset. Additionally, on the scalability test, our augmentation method displays enhanced stability in predicting values that are not included in the training dataset.

2. Method

In this section, we introduce our approach called LAVe, which comprises a labeling, augmentation, and verification model. We denote the conversation history as d_t = (u_1, s_1, u_2, s_2, ..., u_t) where u_t is a user utterance and s_t is a system utterance at turn t. A turn-level belief state at turn t is b_t which consists of slot-value pairs that are mentioned at turn t. Accumulated belief state $B_t = \{b_1, b_2, ..., b_t\}$ represents belief states from turn 1 to t and having all slot-value pairs that user mentioned through the dialogue context. Figure 2 shows an overview of LAVe.

2.1. Teacher Model (Labeling)

The task of the teacher model, also called the labeling model, is to predict the belief states B_t, using the dialogue history d_t and the prompt *"translate dialogue to belief state"* The teacher model is trained with original labeled data $D_{\text{label}} = \{(B_t^n, d_t^n)\}_{n=1,t=1}^{N,T}$ where the N is the number of labeled dialogues. The model is optimized by minimizing the negative log-likelihood of the gold label, which is

$$L_{\text{label}} = -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P(B_t^n|d_t^n). \tag{1}$$

After training this model, we utilize the trained teacher model to generate pseudo-label $\tilde{B}_t$ given d_t from $D_{\text{unlabeled}} = \{d_t^m\}_{m=N+1,t=1}^{M,T}$ where the M is the number of entire dialogues. The teacher model build the pseudo-labeled dataset $D_{\text{pseudo}} = \{(\tilde{B}_t^m, d_t^m)\}_{m=N+1,t=1}^{M,T}$ by combining d_t with $\tilde{B}_t$.

2.2. Augmentation Model

To make the student model understand diverse user utterances, we train the augmentation model that uses D_{label} dataset with prompt *"generate user utterance"* to generate u_t^n given the system utterance s_{t-1}^n and turn-level belief state b_t^n. The loss function is as follows[1].

$$L_{\text{aug}} = -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P(u_t^n|s_{t-1}^n, b_t^n). \tag{2}$$

Then, we create the D_{aug} based on both D_{label} and D_{pseudo} by augmenting user utterance and its corresponding annotation. To make D_{aug}, we first build a value dictionary for each slot by collecting the values from the $B_t^n \in D_{\text{label}}$ and $\tilde{B}_t^m \in D_{\text{pseudo}}$. Using this value dictionary, we augmented the belief state by substituting it with different values. Once the belief states are changed, the augmentation model generates new user utterances that are conditioned by the system utterance and augmented belief state. The augmented dataset is denoted as $D_{\text{aug}} = \text{Aug}(D_{\text{pseudo}} \cup D_{\text{label}})$. We provide an example of this process in Figure 3.

2.3. Verification Model

The verification model aims to distinguish the validity of the dialogue and annotation pair. To train this model, we pair the user and system utterance with its corresponding turn-level belief state $c_t^n = \{s_{t-1}^n, u_t^n, b_t^n\}$ as a positive sample and make negative samples $c_t^{n-} = \{s_{t-1}^n, u_t^n, b_t^{n-}\}$ by modifying the belief state b_t. During modification, we use three functions; (1) addition: adding the random slot and value to the b_t^n, (2) deletion: removing the random slot-value pair from the b_t^n, and (3) substitution: replacing the value in b_t^n. Given the dialogue-belief state pair and prompt *"verify the dialogue-belief pair"* the model generates 'true' or 'false' according to the validness of the pair. Equation 3 is the loss used for the verification model.

$$L_{\text{ver}} = -\frac{1}{2NT}\sum_{n=1}^{N}\sum_{t=1}^{T}[\log(P(\text{true}|c_t^n)) + \log(P(\text{false}|c_t^{n-}))] \tag{3}$$

After we train the verification model, we filter out the incorrect values from D_{pseudo} and D_{aug}, which is simply denoted as $\text{Ver}(D_{\text{peseudo}} \cup D_{\text{aug}})$.

2.4. LAVe and Student Model

The ultimate goal of our research is to develop a student model that surpasses the performance of its teacher model. To accomplish this, we train the student model with the $\text{Ver}(D_{\text{peseudo}} \cup D_{\text{aug}})$, which is verified pseudo-labeled and augmented dataset. The loss function and training method of the student model remain identical to those of the teacher model.

2.5. Implementation Detail

The LAVe is trained based on the pre-trained t5-small [23] and each loss is update by the AdamW [24] optimize function. Models are trained using max to 30 epochs and a learning rate of 1e-3 with a batch size of 16. Training is conducted in A5000x2 GPU with 3 different seeds.

3. Experiment

3.1. Dataset and Metric

MultiWOZ To check the effectiveness of our LAVe framework, we conduct experiments with the MultiWOZ2.0 dataset [2]. It contains approximately 10,000 multi-turn dialogues, which covers seven diverse domains[2]. As we assume the environment where only a few labeled data are available, and the remaining are unlabeled raw dialogue, we make the few-shot environment by splitting the 10% data as labeled and the remaining 90% as unlabeled by deleting the annotated information. For reliable experiment results, we made 3 different 10% labaeld datasets and report the average of the results. For the metric, we use joint goal accuracy (JGA), which is the ratio of correct turns to total turns, and the turn is correct if all of its predicted slot-value pairs (B_t) are equal to the label.

Test Set for Scalability In real-world circumstances, new values are constantly emerging, and DST needs to be robust to these values. However, the current MultiWOZ test set does not reflect this characteristic, as most values in the test set are

[1] In MultiWOZ, the user always starts first, so we left s_0 as empty.

[2] Hotel, Restaurant, Attraction, Train, Taxi, Hospital, and Police

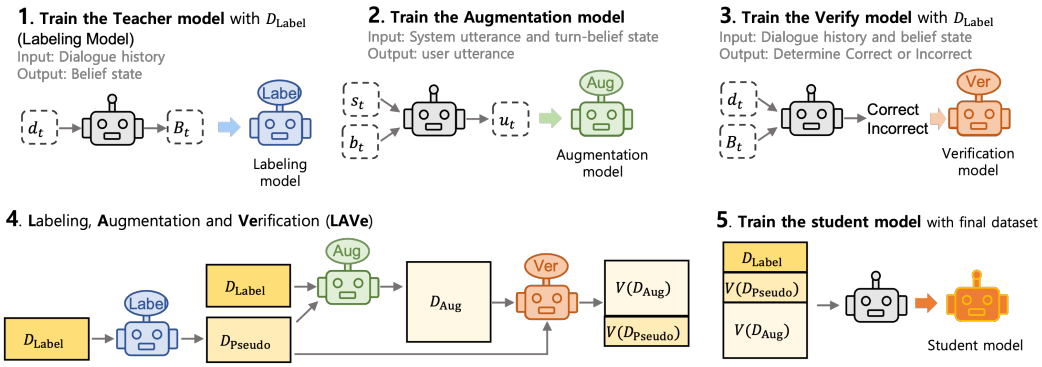

Figure 2: *Overview of LAVe*

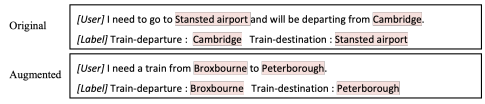

Figure 3: *Example of augmentation result.*

already present in the training dataset. To address this limitation, we curated a new test set that includes diverse unseen values. In creating this new dataset, we replaced the original values with data from different countries, such as Australia and Canada. For instance, we modified the original dialogue "I need to go to Cambridge" with the annotation "taxi-destination: Cambridge" to "I need to go to Broxbourne" with the annotation 'taxi-destination: Broxbourne". Through this process, we created Austrailia and Canada versions MultiWOZ test set, which contains 8409 unseen values .

3.2. Comparison with State-of-the-Art Approaches

Table 1: *Result with the DST models on MWOZ2.0.*

model	JGA	
	10%	100%
SUMBT [3]	-	46.65
TRADE [25]	34.07	48.62
TRADE$_{ssup}$ [26]	37.16	48.72
TOD-BERT [27]	38.80	-
COMER [28]	-	48.79
MINTL [29]	30.32	52.10
PPTOD [30]	45.96	53.89
DS2 [31]	**47.61**	**54.78**
Teacher (Baseline)	46.58 $\pm$ 0.33	54.50
Teacher + LAVe (Student)	**49.75**$\pm$ 0.38	N/A

In this study, we compare the performance of LAVe, on the MultiWOZ2.0 test dataset with other state-of-the-art methods. As our method utilizes 10% of labeled data and 90% of unlabeled raw dialogue, we conduct comparisons in few-shot (10%) as well as fully labeled dataset (100%) scenarios for a fair comparison. We note that LAVe is designed to leverage the benefits of unlabeled data and hence, we do not evaluate its performance in a fully labeled dataset scenario. In Table 1, our findings reveal that LAVe significantly increases the baseline teacher performance in a few-shot learning scenario (46.58% $\rightarrow$ 49.57%). Furthermore, LAVe shows a comparable result with those using fully labeled datasets, even when utilizing only 10% of gold labels. This indicates that LAVe is a powerful technique for leveraging unannotated dialogue.

3.3. Comparison with Confidence Based Selection

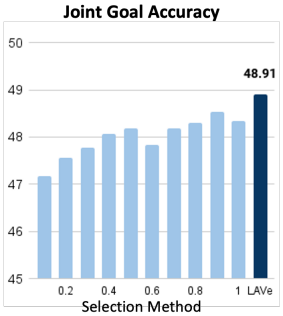

Figure 4: *JGA of student model with different selection methods.*

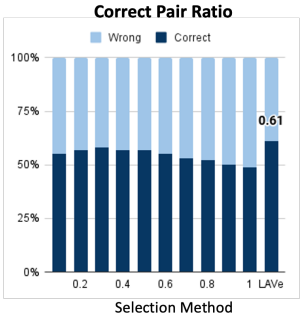

Figure 5: *Ratio of correct dialogue-annotation pairs in selected dataset.*

In this study, we evaluate the effectiveness of our proposed verification model in the LAVe framework by comparing it with the confidence score-based threshold method [11, 12, 13]. The confidence score is calculated by taking the average likelihood of the teacher model for each pseudo label, and we select the top n% confidence-scored pseudo label. The x-axis in Figure 4 and 5 denotes the selection method. For example, x-axis 0.2 means we select top 20% of confident pseudo-label.

JGA of student model In order to assess the validity of the selected dataset, we trained a student model using the selected dataset and compared its JGA against confidence-based model. As illustrated in Figure 4, our proposed method outperforms the confidence-based model, achieving the highest JGA. Intriguingly, we observe that utilizing the entire pseudo-labeled dataset without any confidence-based filtering yields also comparable result. This suggests that although the dataset contains some incorrect pseudo-labels, showing diverse data can increase the student model's accuracy.

How many correct values are in the selected dataset? The dataset's validity is implicitly evaluated by comparing the student model's JGA in Figure 4. In Figure 5, we conduct a more direct analysis of the selected data's quality by illustrating

Table 2: *Results of scalability test with Australia and Canada version test set.*

Augmentation	JGA [%]	
	Australia	Canada
Teacher (Baseline)	33.27± 5.78	34.58± 3.97
EDA	30.32± 4.78	30.38± 4.88
Back Translation	33.45± 3.96	34.05± 4.68
LAVe$_{\text{maintain}}$	33.53± 4.59	35.96± 2.20
AEDA	33.83± 4.13	35.30 ±3.88
LAVe	**36.66** ± 2.64	**36.77** ± 2.46

the ratio of valid dialogue-label pairs within the dataset. This analysis is conducted by comparing the selected dataset with the gold labels and marking the pairs that matched with the gold label as correct and those that did not as incorrect. The results in Figure 5 show that our selected dataset has the highest ratio of correct pairs, indicating its superiority over confidence-based methods. It is worth noting that the trends observed in JGA in Figure 4 and the validity ratio trends in Figure 5 are not aligned. This finding suggests that finding the optimal confidence threshold does not guarantee a high level of validity in the selected dataset. However, our proposed method ensures both a high JGA performance for the student model and a valid dataset selection.

3.4. Scalability Test

To evaluate the generalizability of our augmentation methods, we conducted an analysis with other augmentation methods including EDA[32], AEDA[33], Back Translation[3][34] and LAVe with augmenting using only b_t. Specifically, we examine the performance of our approach using the performance on the Australia and Canada variants of the MultiWOZ test set, which contain unseen values. In table 2, LAVe outperforms other methods by a substantial margin on both datasets. This suggests that our method is highly effective and helps the baseline to be robust to the values that are not present in the training set which is a crucial ability for DST models when they are deployed in the real world.

4. Analysis

4.1. Ablation Study

Table 3: *Ablation study results of the LAVe.*

model	JGA
Teacher (Baseline)	46.58 ± 0.33
+ D_{Pseudo}	48.34 ± 0.43
+ $D_{\text{Pseudo}} \cup D_{\text{Aug}}$	48.48 ± 0.67
+ Ver($D_{\text{Pseudo}} \cup D_{\text{Aug}}$) (LAVe)	**49.75** ± 0.38

To analyze the effects of our labeling, augmentation, and verification methods, we conduct an ablation study by adding each method to the baseline (Table 3). In the table, notation Ver($D_{\text{pseudo}} \cup D_{\text{aug}}$) means verified result of D_{pseudo} and D_{aug} dataset. Overall, our method increases the baseline by 3.17% by leveraging the unlabeled raw dialogue. More specifically, our pseudo-labeling method increased the accuracy by 1.76%, and with combined augmentation, the difference increased to 1.9%. Furthermore, by applying the verification method to $D_{\text{pseudo}} \cup D_{\text{aug}}$, the accuracy increases 1.27%, additionally. This result shows that our labeling, augmentation and verification models all help to increase the baseline, and the verification model effectively filter out the incorrect value from D_{pseudo} and D_{aug}.

[3]Helsinki-NLP/opus-mt-en-de in https://huggingface.co/

Table 4: *Comparison of error rate per each type with the baseline. The numer in parenthesis are actual amount of errors.*

model	Error Type		
	Wrong	Ignore	Spurious
Teacher (Baseline)	▽0% (3219)	▽0% (2553)	▽0% (3072)
+ D_{pseudo}	▽2.54% (3138)	▽7.12% (2372)	▽11.79% (2709)
+ $D_{\text{pseudo}} \cup D_{\text{aug}}$	▽2.55% (3137)	▽5.89% (2403)	▽**13.91**% (2644)
+ Ver($D_{\text{pseudo}} \cup D_{\text{aug}}$)	▽2.74% (3131)	▽**10.46**% (2287)	▽13.95% (2643)

4.2. Error Analysis

To identify which part of our method reduces the error, we conduct ablation studies and check the decreased rate for each error type. Table 4 presents the percentage of reduced error rate compared to the baseline, and the scripted number in parentheses indicates the actual error cases. The type "Wrong" indicates the model predicts a different value despite being correct in predicting a slot. In contrast, "Ignore" means the model predicts no value for a slot that is mentioned in dialogue. Finally, "Spurious" denotes the model predicted slot not mentioned in dialogue. The experiment results show that pseudo-labeling effectively reduces all types of errors; this shows that our labeling model effectively utilizes the raw dialogue for students. However, although the augmentation model decreases spurious errors, it leads to a slight increase in ignore-type errors. This indicates that the augmentation model may generate some erroneous dialogue-label pairs that do not accurately capture the augmented label information. Nonetheless, this issue is mitigated by our verification model, which filters out incorrect pairs and ultimately reduces overall error rates.

5. Conclusion

In this study, we introduce a new framework, LAVe, for self-training in DST. LAVe includes a reliable verification method to ensure the student model's accuracy and the dataset's validity. Additionally, to effectively utilize the pseudo-labeled dataset, we propose an augmentation method that modifies the pseudo-label. In the experiment, even when using 10% of the labeled dataset, LAVe performs similarly to those trained on fully labeled datasets and shows scalability to the value that is not in the training sets. We expect LAVe to be a good reference to leveraging raw dialogues in dialogue-related research.

Acknowledgements: This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) (No.2021-0-00575, Development of Voicepishing Prevention Technology Based on Speech and Text Deep Learning), IITP (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)) grant funded by the Korea government (MSIT) and, MSIT support program (IITP-2023-2020-0-01789) supervised by the IITP.

6. References

[1] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[2] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," *arXiv preprint arXiv:1810.00278*, 2018.

[3] H. Lee, J. Lee, and T.-Y. Kim, "Sumbt: Slot-utterance matching for universal and scalable belief tracking," *arXiv preprint arXiv:1907.07421*, 2019.

[4] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 179–20 191, 2020.

[5] M. Heck, C. van Niekerk, N. Lubis, C. Geishauser, H.-C. Lin, M. Moresi, and M. Gašić, "Trippy: A triple copy strategy for value independent neural dialog state tracking," *arXiv preprint arXiv:2005.02877*, 2020.

[6] D. Ham, J.-G. Lee, Y. Jang, and K.-E. Kim, "End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 583–592.

[7] Y. Yang, Y. Li, and X. Quan, "Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 230–14 238.

[8] Z. Zhang, J. Han, J. Deng, X. Xu, F. Ringeval, and B. Schuller, "Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning," *IEEE Access*, vol. 6, pp. 22 196–22 209, 2018.

[9] X. Li, Q. Sun, Y. Liu, Q. Zhou, S. Zheng, T.-S. Chua, and B. Schiele, "Learning to self-train for semi-supervised few-shot classification," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bf25356fd2a6e038f1a3a59c26687e80-Paper.pdf

[10] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, "Self-training for end-to-end speech translation," *arXiv preprint arXiv:2006.02490*, 2020.

[11] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020.

[12] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[13] Y. Xu, L. Shang, J. Ye, Q. Qian, Y.-F. Li, B. Sun, H. Li, and R. Jin, "Dash: Semi-supervised learning with dynamic thresholding," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 525–11 536.

[14] D. Berthelot, R. Roelofs, K. Sohn, N. Carlini, and A. Kurakin, "Adamatch: A unified approach to semi-supervised learning and domain adaptation," *arXiv preprint arXiv:2106.04732*, 2021.

[15] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the ninth international conference on Information and knowledge management*, 2000, pp. 86–93.

[16] T. Vu, M.-T. Luong, Q. V. Le, G. Simon, and M. Iyyer, "Strata: Self-training with task augmentation for better few-shot learning," *arXiv preprint arXiv:2109.06270*, 2021.

[17] F. Mi, W. Zhou, F. Cai, L. Kong, M. Huang, and B. Faltings, "Self-training improves pre-training for few-shot learning in task-oriented dialog systems," *arXiv preprint arXiv:2108.12589*, 2021.

[18] X. Xu, G. Wang, Y.-B. Kim, and S. Lee, "Augnlg: Few-shot natural language generation using self-trained data augmentation," *arXiv preprint arXiv:2106.05589*, 2021.

[19] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] J. He, J. Gu, J. Shen, and M. Ranzato, "Revisiting self-training for neural sequence generation," *arXiv preprint arXiv:1909.13788*, 2019.

[21] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7084–7088.

[22] K. Veselý, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 267–272.

[23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[25] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung, "Transferable multi-domain state generator for task-oriented dialogue systems," *arXiv preprint arXiv:1905.08743*, 2019.

[26] C.-S. Wu, S. Hoi, and C. Xiong, "Improving limited labeled dialogue state tracking with self-supervision," *arXiv preprint arXiv:2010.13920*, 2020.

[27] C.-S. Wu, S. Hoi, R. Socher, and C. Xiong, "Tod-bert: Pre-trained natural language understanding for task-oriented dialogue," *arXiv preprint arXiv:2004.06871*, 2020.

[28] L. Ren, J. Ni, and J. McAuley, "Scalable and accurate dialogue state tracking via hierarchical sequence generation," *arXiv preprint arXiv:1909.00754*, 2019.

[29] Z. Lin, A. Madotto, G. I. Winata, and P. Fung, "Mintl: Minimalist transfer learning for task-oriented dialogue systems," *arXiv preprint arXiv:2009.12005*, 2020.

[30] Y. Su, L. Shu, E. Mansimov, A. Gupta, D. Cai, Y.-A. Lai, and Y. Zhang, "Multi-task pre-training for plug-and-play task-oriented dialogue system," *arXiv preprint arXiv:2109.14739*, 2021.

[31] J. Shin, H. Yu, H. Moon, A. Madotto, and J. Park, "Dialogue summaries as dialogue states (ds2), template-guided summarization for few-shot dialogue state tracking," *arXiv preprint arXiv:2203.01552*, 2022.

[32] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[33] A. Karimi, L. Rossi, and A. Prati, "Aeda: An easier data augmentation technique for text classification," *arXiv preprint arXiv:2108.13230*, 2021.

[34] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.