Semi-Supervised Lifelong Language Learning

Yingxiu Zhao[1]*, **Yinhe Zheng**[2]†, **Bowen Yu**[2], **Zhiliang Tian**[1], **Dongkyu Lee**[1], **Jian Sun**[2], **Yongbin Li**[2]†, **Nevin L. Zhang**[1]

[1] The Hong Kong University of Science and Technology, Hong Kong SAR, China

[2] Alibaba Group, China

{yzhaocx,ztianac,dleear,lzhang}@connect.ust.hk, zhengyinhe1@163.com,

{yubowen.ybw,shuide.lyb}@alibaba-inc.com, jiansun_china@hotmail.com

Abstract

Lifelong learning aims to accumulate knowledge and alleviate catastrophic forgetting when learning tasks sequentially. However, existing lifelong language learning methods only focus on the supervised learning setting. Unlabeled data, which can be easily accessed in real-world scenarios, are underexplored. In this paper, we explore a novel setting, semi-supervised lifelong language learning (SSLL), where a model learns sequentially arriving language tasks with both labeled and unlabeled data. We propose an unlabeled data enhanced lifelong learner to explore SSLL. Specially, we dedicate task-specific modules to alleviate catastrophic forgetting and design two modules to exploit unlabeled data: (1) a virtual supervision enhanced task solver is constructed on a teacher-student framework to mine the underlying knowledge from unlabeled data; and (2) a backward augmented learner is built to encourage knowledge transfer from newly arrived unlabeled data to previous tasks. Experimental results on various language tasks demonstrate our model's effectiveness and superiority over competitive baselines under the new setting SSLL. We will release the code and data [1].

1 Introduction

A remarkable ability of humans is to learn and accumulate knowledge continuously throughout their lifetime. Such *Lifelong Learning* ability is crucial for computational systems interacting with the real world and processing continuous streams of information (Parisi et al., 2019; Delange et al., 2021). However, most deep neural networks studies assume data distributions are stationary, which is not applicable in the real-world environments that dynamically evolve. In such real scenarios, models often suffer from *catastrophic forgetting* (McClelland et al., 1995; Parisi et al., 2019): a phenomenon where models forget previously learned knowledge when learning new tasks sequentially.

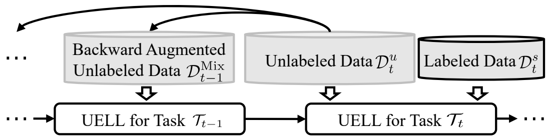

Figure 1: The Training process of our model UELL.

Various approaches have been proposed to alleviate catastrophic forgetting in lifelong scenarios. Attempts include constraining the variants of important weights with regularization (Schwarz et al., 2018; Mi et al., 2020), storing real samples or using pseudo samples for previous tasks to maintain the learned knowledge (d'Autume et al., 2019; Sun et al., 2019; Chuang et al., 2020), or dedicating task-specific modules to avoid the interference among tasks (Madotto et al., 2021; Qin and Joty, 2022; Zhang et al., 2022). Despite their reported effectiveness, these approaches are mostly designed to handle supervised learning tasks, where only labeled data are available. In real-world scenarios, labeled data are generally expensive and time-consuming to obtain, whereas unlabeled data are much easier to collect. These unlabeled data often carry rich information and have been successfully utilized to improve model performance in semi-supervised learning (Xie et al., 2020a; Chen et al., 2021).

In this paper, we investigate a novel setting: Semi-Supervised Lifelong Language learning (SSLL), where a model learns sequentially arriving language tasks with limited labeled data and adequate unlabeled data (see the training process in Fig. 1). The abundant information in the unlabeled data can not only facilitate learning the current task but also benefit learned tasks with similar data distributions. For example, sentiment analysis and topic classification tasks may only differ in their la-

* Work done while the author was interning at Alibaba.

† Corresponding author.

[1] https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/ssll

bel spaces while sharing the same set of unlabeled data. We can transfer knowledge among the two kinds of tasks using these unlabeled data (Liu et al., 2019). This phenomenon naturally leads to two challenges to be faced in the SSLL scenario: **(1)** ***How to fully exploit unlabeled data to facilitate each arrived language task?*** and **(2)** ***How to leverage newly arrived unlabeled data to encourage knowledge transfer to previous tasks?***

With this in mind, we propose an Unlabeled data Enhanced Lifelong Learner (UELL) framework to explore the new setting SSLL. Specifically, we dedicate task-specific parameters to alleviate catastrophic forgetting in UELL. We construct two modules to tackle the challenges mentioned above. The first module is a virtual supervision enhanced solver that exploits unlabeled data using a teacher-student framework. The teacher generates pseudo labels for unlabeled data as virtual supervision and guides the student according to its learning progress. The student also learns from pseudo labels through self-study. The second module is a backward augmented learner that encourages knowledge transfer from the current task to previously learned tasks. The generated pseudo data for each learned task are augmented by retrieving semantically similar unlabeled samples from the current task, where the latter are leveraged to transfer knowledge backward to previously learned tasks. We conduct extensive experiments and analyses on both language understanding and generation tasks and demonstrate that UELL can effectively address the challenges of SSLL.

Our main contributions are as follows:

- To the best of our knowledge, we are the first to explore the semi-supervised lifelong language learning setting, where a model learns sequentially arriving language tasks with a mixture of labeled and unlabeled data.
- We propose a novel method, Unlabeled data Enhanced Lifelong Learner, to exploit unlabeled data and encourage backward knowledge transfer in SSLL.
- We conduct adequate experiments and analyses on various language tasks. The results demonstrate the effectiveness of UELL over competitive baselines adapted from lifelong learning methods to the SSLL setting.
- We believe our new paradigm imposes new challenges and opens up new research opportunities for the NLP community.

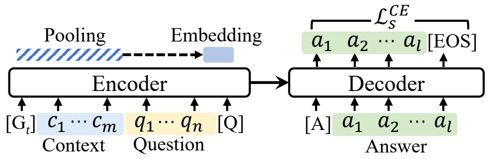

Figure 2: The format of UELL's input-output. "[G_t]" is the task-specific generation token for task $\mathcal{T}_t$, "[Q], [A]" are special tokens that indicate the end of the question and the beginning of the answer, respectively.

2 Related Work

Semi-Supervised Learning aims to learn from both labeled and unlabeled data (Chapelle et al., 2009; Van Engelen and Hoos, 2020; He et al., 2022a). As an efficient approach for semi-supervised learning, pseudo labeling (Zhou and Li, 2010; Lee et al., 2013) tries to utilize unlabeled data by predicting their labels. Some studies (Zhou and Li, 2010; Qiao et al., 2018) train multiple learners and exploit disagreements among different learners. Other studies utilize self-training to generate pseudo labels for unlabeled data (Zhai et al., 2019; Xie et al., 2020b; Chen et al., 2020).

Consistency regularization (Rasmus et al., 2015; Tarvainen and Valpola, 2017; Xie et al., 2020a; Wu et al., 2021; He et al., 2022c) is another popular scheme. It regularizes the model to be invariant to small perturbations on the input, hidden states, or model parameters. Mean Teacher (Tarvainen and Valpola, 2017) is an efficient method to implement consistency regularization, where a student model and a teacher model are maintained to enforce the predictions' consistency. This paper combines pseudo labeling with a teacher-student framework to handle semi-supervised learning in the lifelong scenarios.

Lifelong Learning aims to learn a sequence of tasks without forgetting previously learned knowledge. Three categories of approaches are generally used in lifelong learning: *Regularization-based* methods either impose constraints on the variation of important weights when learning new tasks (Schwarz et al., 2018; Aljundi et al., 2018), or introduce knowledge distillation to preserve the previously learned knowledge (Li and Hoiem, 2017; Dhar et al., 2019); *Replay-based* methods store real samples (Rebuffi et al., 2017) or generate pseudo samples (Qin and Joty, 2022; Zhao et al., 2022b) for learned tasks to consolidate previous knowledge; *Architecture-based* methods construct task-

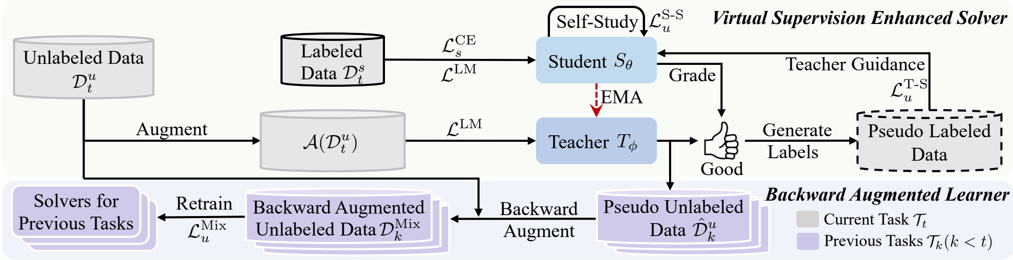

Figure 3: Overview of our method UELL. UELL consists of two modules: a virtual supervision enhanced solver to exploit unlabeled data and a backward augmented learner to encourage knowledge transfer to previous tasks.

specific modules to preserve knowledge. Some studies (Serrà et al., 2018; Fernando et al., 2017) use static architectures and apply task-specific routes through the architectures to prevent forgetting, while other studies (Madotto et al., 2021; Zhang et al., 2022; Dai et al., 2022) dynamically expand the model with task-specific parameters.

Recently, some studies have tried to investigate semi-supervised lifelong learning for image classification tasks by modeling data distributions with generative adversarial networks (Goodfellow et al., 2014), or relying on a super-class structure of image datasets to exploit unlabeled data (Wang et al., 2021; Brahma et al., 2021; Smith et al., 2021). However, it is non-trivial to extend these works into language tasks since they either require continuous input spaces or utilize extra super-class structures, and all these works do not consider backward knowledge transfer with the help of unlabeled data.

3 Task Definition and Formulation

In SSLL, we sequentially learn a stream of semi-supervised language tasks $\mathcal{T}_1$, $\mathcal{T}_2$, ..., $\mathcal{T}_N$. The data of each task $\mathcal{T}_t$ contain limited labeled data $\mathcal{D}_t^s = \{(X_i^s, Y_i^s)\}$ and abundant unlabeled data $\mathcal{D}_t^u = \{X_i^u\}$, where X_i^s and X_i^u are input samples and Y_i^s is the output label of X_i^s. Moreover, the data for task $\mathcal{T}_t$ arrive after task $\mathcal{T}_{t-1}$ is learned (Brahma et al., 2021), and we have no access to previous tasks' data. The final goal is to optimize the model's average performance on **all tasks** after training the whole sequence (Sun et al., 2019).

Inspired by McCann et al. (2018), we frame different types of NLP tasks into a unified text-to-text format. Specifically, for each $(X^s, Y^s) \in \mathcal{D}^s$, we format X^s as a concatenation of a context and a question, and serialize Y^s as an answer sequence representing the label of X^s (Fig. 2). For instance, a sample in a sentiment classification task contains the input X^s: "`I enjoyed the movie.` (context) `What's the sentiment?` (question)" and the output Y^s: "`positive`". For each sample $X^u \in \mathcal{D}^u$, the input X^u is only a context.

4 Methodology

4.1 Overview

We propose an unlabeled data enhanced lifelong learner (UELL) to handle the newly proposed semi-supervised lifelong language learning (SSLL) setting (see Fig. 1 and 3). To alleviate the catastrophic forgetting issue of lifelong learning, we dynamically expand the model architecture by allocating task-specific modules. To overcome the challenges of SSLL discussed in §1, we design two modules in UELL: (1) a virtual supervision enhanced solver exploits unlabeled data for each sequentially arrived task; (2) a backward augmented learner encourages knowledge transfer to previous tasks with unlabeled data. We will first illustrate the architecture we design to prevent forgetting and then elaborate on details about our learning process.

Model Architecture UELL maintains a teacher model T_ϕ and a student model S_θ for each task during the lifelong learning process. It uses T5 (Raffel et al., 2020) as the backbone for T_ϕ and S_θ to tackle text-to-text generation tasks. To overcome catastrophic forgetting, we freeze the parameters of the pre-trained T5 and insert separate adapter modules (Houlsby et al., 2019; Pfeiffer et al., 2020b) into T5 for each task (See Fig 4). Specifically, we use the bottle-necked adapter structure proposed by Pfeiffer et al. (2020a), in which the adapter consists of a layer normalization (Ba et al., 2016) followed by a two-layer MLP and a residual connection (He et al., 2016). Such the adapters are light-weighted (i.e., 0.8% of T5 parameters). By encapsulating task-specific information into isolated parameters, these adapters can effectively avoid catastrophic forgetting with little memory overhead.

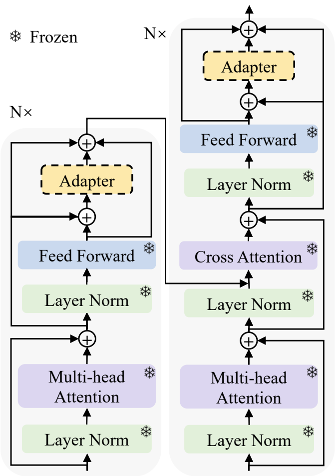

Figure 4: Adapter modules for T5 transformers layers used in UELL. The parameters of the pre-trained T5 (i.e., layers marked with ❄) are fixed during training.

Labeled Data Learning We optimize the student model S_θ over $\mathcal{D}_t^s$ for each arriving task $\mathcal{T}_t$ with the following cross-entropy loss (see Fig. 2):

$$\mathcal{L}_s^{\text{CE}} = -\sum_{(X,Y)\in\mathcal{D}_t^s} \log S_\theta(Y|X), \tag{1}$$

The teacher T_ϕ is gradually updated by means of momentum according to the student's weights θ. We will elaborate on it in §4.4.

4.2 Virtual Supervision Enhanced Solver

To fully exploit unlabeled data for each task, a virtual supervision enhanced solver is constructed with a teacher-student framework: (1) The teacher T_ϕ predicts pseudo labels for unlabeled data and uses these virtual supervisory signals to guide the student; (2) The student S_θ takes a self-study course with virtual signals to further enhance itself.

Teacher Guidance Inspired by (Tarvainen and Valpola, 2017), the guidance from the teacher T_ϕ to the student S_θ is achieved by forcing the consistency between predictions of T_ϕ and S_θ on perturbed unlabeled inputs. Specifically, when learning the task $\mathcal{T}_t$, we augment each unlabeled sample $X^u \in \mathcal{D}_t^u$ to $\mathcal{A}(X^u)$ as the input of the teacher, where $\mathcal{A}$ refers to the data augmentation operation that injects noise to the input while preserving its semantics. Next, the teacher T_ϕ predicts the pseudo label $\hat{Y}$ of $\mathcal{A}(X^u)$ through greedy decoding. To alleviate noises introduced by pseudo labels, we only maintain high confident predictions produced by T_ϕ to guide S_θ. Following Madotto et al. (2021), we use the perplexity score $\text{PPL}(\hat{Y})$ to measure T_ϕ's confidence when predicting $\hat{Y}$. Here, low perplexity corresponds to high confidence of $\hat{Y}$. We optimize S_θ on the filtered pseudo labels using:

$$\mathcal{L}_u^{\text{T-S}}(\mathcal{D}_t^u) = -\sum_{X^u\in\mathcal{D}_t^u} \mathbb{1}(\text{PPL}(\hat{Y}) \leq \tau) \cdot \log S_\theta(\hat{Y}|\mathcal{A}(X^u)), \tag{2}$$

where τ is a confidence threshold.

Moreover, to better guide the student S_θ, we argue that the teacher T_ϕ should adjust its teaching pace based on the learning progress of S_θ. Inspired by Pham et al. (2021), we use the training loss of S_θ on labeled data as its grade to inform the teacher. Considering that unlabeled data are usually "harder" to learn than labeled data, we teach S_θ with unlabeled data only when the student acquires a certain grade (i.e., knowledgeable enough). Specifically, the optimization of $\mathcal{L}_u^{\text{T-S}}$ is carried out only when the absolute difference of losses between S_θ and T_ϕ on labeled data drops to a certain threshold γ.

Student Self-Study In addition to the teacher's guidance, self-study is also crucial for the student to better leverage unlabeled data. The self-study course taken by the student is performed by enforcing the student's consistency under small disturbances using a dropout-based regularization term (Wu et al., 2021). Specifically, for each pseudo labeled data $(\mathcal{A}(X^u), \hat{Y})$, we forward $\mathcal{A}(X^u)$ twice through the student model S_θ with different dropout masks to obtain two different predicted outputs: $S_\theta^i(\hat{Y}|\mathcal{A}(X^u))$, $(i = 1, 2)$. Denote w_k as the k-th token in $\hat{Y}$, $S_\theta^i(w_k)$ as the predicted distribution for token w_k in the i-th forward pass of S_θ. We optimize the following bidirectional Kullback-Leibler (KL) divergence:

$$\begin{aligned}\mathcal{L}_u^{\text{S-S}}(\mathcal{D}_t^u) &= \frac{1}{2}\sum_{X^u\in\mathcal{D}_t^u} KL^{12}(X^u) + KL^{21}(X^u),\\ KL^{ij}(X^u) &= \frac{1}{|\hat{Y}|}\sum_{k=1}^{|\hat{Y}|} KL(S_\theta^i(w_k)||S_\theta^j(w_k)),\end{aligned} \tag{3}$$

where KL measures the KL divergence between two distributions.

4.3 Backward Augmented Learner

Besides enhancing the current task with virtual supervision on unlabeled data, we find that unlabeled data can be used to encourage backward knowledge transfer. We build a backward augmented lifelong learner to leverage the newly arrived unlabeled data to improve solvers for previously learned tasks. This scheme contains two steps: (1) acquiring unlabeled data for learned tasks (§4.3.1), and (2) retraining previous solvers (§4.3.2).

4.3.1 Previous Unlabeled Data Acquisition

Considering that unlabeled data usually contain rich semantic information, the knowledge of subsequent tasks may facilitate previously learned tasks. Specifically, when the unlabeled data $\mathcal{D}_k^u$ of the previous task $\mathcal{T}_k$ $(k < t)$ share similar distributions with the current task $\mathcal{T}_t$, we can augment $\mathcal{D}_k^u$ by retrieving similar samples from the unlabeled data $\mathcal{D}_t^u$ of the currently arrived task $\mathcal{T}_t$ based on samples in $\mathcal{D}_k^u$. However, it is non-trivial to implement the above augmentation process in SSLL because $\mathcal{D}_k^u$ $(k < t)$ are unavailable when learning $\mathcal{T}_t$ (we have no access to previous data).

To tackle the above issue, we equip UELL with the ability to generate pseudo unlabeled data that obey the distribution of $\mathcal{D}_k^u$. In this way, we can retrieve $\mathcal{D}_t^u$ using pseudo unlabeled samples of previous tasks to achieve the aforementioned backward augmentation. This data generation process is optimized through the language modeling loss on both labeled and unlabeled data as follows,

$$\mathcal{L}^{\text{LM}} = -\sum_{X \in \mathcal{D}_i^s} \log S_\theta(X|G_t) - \mu \sum_{X \in \mathcal{D}_i^u} \log S_\theta(X|G_t), \quad (4)$$

where G_t is a task-specific generation token for task $\mathcal{T}_t$, and μ is the weight for the unlabeled loss. $\mathcal{L}^{\text{LM}}$ is only optimized on the context tokens of X. Then, by feeding the generation token G_k to the encoder of T5, we can sample pseudo unlabeled data of the previous task $\mathcal{T}_k$ from the decoder with the top-K sampling scheme.

Note that UELL does not further predict labels for the generated unlabeled data to avoid accumulating errors introduced by noisy pseudo-labels. Moreover, the generated pseudo data are not aimed to prevent forgetting because task-specific adapter modules in UELL are efficient enough to avoid interference among tasks.

4.3.2 Previous Solvers Retraining

When learning the current task $\mathcal{T}_t$, UELL first generates a set of pseudo samples $\{\hat{\mathcal{D}}_k^u\}_{k=1}^{t-1}$ for all previously learned tasks $\{\mathcal{T}_k\}_{k=1}^{t-1}$. To achieve the backward augmentation for each learned task, we utilize samples in $\hat{\mathcal{D}}_k^u$ to retrieve semantically similar unlabeled data in $\mathcal{D}_t^u$ of task $\mathcal{T}_t$. Specifically, UELL encodes the sample contexts into representations through the T5 encoder [2] with an average pooling layer (see Fig. 2). Cosine similarities of

[2]Here, we use the fixed pre-trained T5 encoder without adapters here to prevent the representations from drifting as new tasks are learned.

Algorithm 1 UELL Training

Input: Semi-supervised tasks $\{\mathcal{T}_t\}_{t=1}^N$, A pretrained T5 model, and randomly initialized student and teacher with parameters θ_0 and ϕ_0, respectively. A learning rate η, EMA decay rate α.
Output: Learned teacher parameters θ_t for tasks $\{\mathcal{T}_t\}_{t=1}^N$

```
for t = 1 to N do
    Initialize student φ_t and teacher θ_t using φ_{t−1}
    while Not Converge do
        Sample batches B^s ⊆ D_t^s and B^u ⊆ D_t^u.
        Compute L_s^CE (Eq.1) and L^LM (Eq.4).
        if student θ_t reaches grade γ then
            Augment B^u to A(B^u).
            Predict Ŷ for each X ∈ A(B^u) using T_{φ_t}.
            Compute L_u^T-S (Eq.2) and L_u^S-S (Eq.3).
        end if
        Compute the total loss L (Eq. 6).
        Update student θ_t ← θ_t − η∇L.
        Update teacher φ_t ← αφ_t + (1 − α)θ_t.
    end while
    for k = 1 to t − 1 do
        Generate pseudo data D̂_k^u for task T_k.
        Backward augment D̂_k^u to D_k^Mix
        Update student θ_k and teacher φ_k with L_u^Mix (Eq. 5)
    end for
end for
```

these representations are used to measure the distance of samples. For each sample $X^u \in \hat{\mathcal{D}}_k^u$, we retrieve K nearest neighbors from $\mathcal{D}_k^u$ and augment $\hat{\mathcal{D}}_k^u$ with the $K \cdot |\hat{\mathcal{D}}_k^u|$ retrieved neighbors to produce a set of backward augmented unlabeled data $\mathcal{D}_k^{\text{Mix}}$. Then $\mathcal{D}_k^{\text{Mix}}$ is used to enhance the learned solver for $\mathcal{T}_k$ by optimizing the losses on Eq. 2 and 3:

$$\mathcal{L}_u^{\text{Mix}}(\mathcal{D}_k^{\text{Mix}}) = \mathcal{L}_u^{\text{T-S}}(\mathcal{D}_k^{\text{Mix}}) + \mathcal{L}_u^{\text{S-S}}(\mathcal{D}_k^{\text{Mix}}). \quad (5)$$

In this way, we can encourage the knowledge transfer from newly encountered unlabeled data to previously learned tasks.

4.4 Model Update Procedure

Before learning the first task, the teacher and student models in UELL are initialized with randomly-initialized adapters layers with the pre-trained T5 backbone. When learning the current task $\mathcal{T}_t$, the student model S_θ is trained using the following loss:

$$\mathcal{L} = \mathcal{L}_s^{\text{CE}} + \mu\mathcal{L}_u^{\text{T-S}}(\mathcal{D}_t^u) + \mu\mathcal{L}_u^{\text{S-S}}(\mathcal{D}_t^u) + \lambda\mathcal{L}^{\text{LM}}, \quad (6)$$

where λ is the weight to balance the task learning and language modeling. To prevent confirmation bias (Tarvainen and Valpola, 2017), the teacher weights ϕ are updated as an exponential moving average (EMA) of student weights in each batch:

$$\phi_p = \alpha\phi_{p-1} + (1 - \alpha)\theta_p, \quad (7)$$

where α is the EMA decay rate, p is the time step. Note that the slowly evolved teacher can be regarded as an ensemble of student models in different training iterations. This leads to more stable and accurate predictions on unlabeled data (Tarvainen and Valpola, 2017). After learning task $\mathcal{T}_t$, UELL generates pseudo unlabeled data for each previously learned tasks $\{\mathcal{T}_k\}_{k=1}^{t-1}$ and further optimizes the learned solvers with the loss $\mathcal{L}_u^{\text{Mix}}(\mathcal{D}_k^{\text{Mix}})$ shown in Eq.5 to enable the backward knowledge transfer. See Algorithm 1 for more details.

5 Experiment Setup

5.1 Datasets

Following Sun et al. (2019), we evaluate our approach from two dimensions: (1) tasks with the Same Type but Different Domains (STDD); (2) tasks of Different Types (DT). For STDD, we follow Sun et al. (2019) to use five text classification datasets covering domains from news classification, sentiment analysis, and Wikipedia article classification. We follow d'Autume et al. (2019) to produce balanced datasets. For DT, we consider five different sequence generation tasks from decaNLP (McCann et al., 2018): question answering, semantic parsing, semantic role labeling, goal-oriented dialogue generation, and sentiment analysis. To each task $\mathcal{T}_t$ in the SSLL setting, we randomly select 100 labeled data to construct $\mathcal{D}_t^s$ and select another 2,000 unlabeled data to construct $\mathcal{D}_t^u$. More details are provided in Table 1 and Appendix A.1.

5.2 Implementation Details

We use T5-base (Raffel et al., 2020) as our backbone and implement adapters using AdapterHub (Pfeiffer et al., 2020b). We set the confidence threshold $\tau = 1.5$ (Eq.2), unlabeled loss weight $\mu = 0.01$ (Eq.4), language modeling loss weight $\lambda = 0.5$ (Eq.6), and EMA decay rate $\alpha = 0.95$ (Eq.7). We set the threshold of teacher guidance γ to 0.1 in §4.2 and choose $K = 3$ nearest neighbors in §4.3. We train our model UELL on 1 Tesla-V100 GPU. Each task in STDD and DT is trained using the Adam optimizer (Kingma and Ba, 2015) for 120 and 200 epochs, respectively, with a warm-up ratio of 0.1 and maximum learning rate of 2e-4. It takes around 5 and 18 hours to learn all STDD tasks and DT tasks, respectively. The training and testing batch size is set to 16. We use EDA (Wei and Zou, 2019) to implement the data augmentation on unlabeled data $\mathcal{D}^u$ as $\mathcal{A}(\mathcal{D}^u)$. For backward augmentation, we train one epoch on the augmented unlabeled data $\mathcal{D}_k^{\text{Mix}}$ for previous task $\mathcal{T}_k$ to optimize the previously learned solver.

All results reported in this paper are averages of five different runs with random task orders.

5.3 Baselines

We compare our model with the following baselines. **Fine-tuning** (FT) directly tunes a pretrained T5 model on incoming data. *Regularization-based methods*: **EWC** (Schwarz et al., 2018) and **MAS** (Aljundi et al., 2018) mitigate forgetting by penalizing variation of important parameters for previous tasks; *Replay-based methods*: **ER** (Rolnick et al., 2019) stores real samples of learned tasks to prevent forgetting. **LAMOL**(Sun et al., 2019) generates pseudo samples of previous tasks and trains them with the new tasks' data; *Architecture-based methods*: **HAT** (Serrà et al., 2018) uses a task-specific hard attention mechanism to preserve previously learned tasks. **CTR** introduces continual learning plugins into BERT and uses task masks to preserve task-specific knowledge and encourage knowledge transfer [3]; **Adapter** (Madotto et al., 2021) dynamically expands the model by assigning task-specific adapters. For fair comparisons, we use a pretrained T5 as its backbone. **Compositional-Adapters** (Comp) (Zhang et al., 2022) utilizes hidden state mixing to adaptively compose old and new adapters for new tasks and employs generative replay to facilitate knowledge transfer. Besides the above baselines, we also test the **Multi-task Learning** (MTL) approach that tunes the whole model to learn all tasks simultaneously. This approach is usually seen as an upper bound of lifelong learning.

Note that all the above baselines focus on supervised lifelong learning. To enable a fair comparison under the SSLL setting, we enhance the baselines with a strong pseudo labeling method to utilize unlabeled data (Xie et al., 2020b). Specifically, for each task, we first train each baseline model on labeled data. We generate pseudo labels for unlabeled data and train the mixture of labeled data and pseudo labeled data as the final model. For compared baselines, we follow their original settings for training. See more details in Appendix A.2.

[3] CTR and HAT are specifically designed for classification tasks, and it is non-trivial to extend them to text-generation tasks, so we do not compare them for DT tasks.

Dimensions	Datasets	Metrics	# Training Set	# Testing Set
STDD	AGNews	EM	115,000	7,600
	Amazon	EM	115,000	7,600
	DBPedia	EM	115,000	7,600
	Yahoo	EM	115,000	7,600
	Yelp	EM	115,000	7,600
DT	SQuAD	nF1	87,599	10,570
	WikiSQL	lfEM	56,355	15,878
	SST	EM	6,920	1,821
	QA-SRL	nF1	6,414	2,201
	WOZ	dsEM	2,536	1,646

Table 1: Summary of datasets statics and their metrics. nF1 is the normalized version of the F1 score; EM represents an exact match between texts: for text classification tasks) this amounts to accuracy; for WOZ, it is equivalent to dfEM (turn-based dialogue state exact match); for WikiSQL, it is equivalent to lfEM (exact match of logical forms).

5.4 Metrics

Following Sun et al. (2019), we evaluate each task with its corresponding metric (see Table 1). The score of each metric lies from 0 to 100%. We evaluate the performance of lifelong learning using *Average Score* (**Avg-Score**) (Sun et al., 2019; Madotto et al., 2021) that measures the average test scores of all learned N tasks:

$$\text{Avg-Score} = \frac{1}{N}\sum_{j=1}^{N} R_{N,j},$$

where $R_{i,j}$ is the test score of task $\mathcal{T}_j$ after the i-th task is learned. Following Lopez-Paz and Ranzato (2017), we evaluate the effect of *backward knowledge transfer* (**BWT**) to assess the impact of learning on subsequent tasks on previously learned tasks:

$$\text{BWT} = \frac{1}{N-1}\sum_{j=1}^{N-1} R_{N,j} - R_{j,j}.$$

A negative BWT indicates that the model has forgotten some previously acquired knowledge, i.e., suffers from catastrophic forgetting. In general, the higher of these two metrics, the better the model.

6 Experimental Results and Analyses

6.1 Main Results

Table 2 and 3 show the performances of all methods. UELL significantly outperforms all baselines by a large margin. We can also observe that:

- Directly tuning a single model sequentially (the FT baseline) suffers from severe catastrophic forgetting issues, highlighting the importance of lifelong learning studies.
- Regularization based methods EWC and MAS improve the lifelong learning performance to some extent, but they still perform inferior to replay-based methods ER and LAMOL.
- ER outperforms LAMOL on the Avg-Score, indicating that real samples carry higher-quality knowledge than pseudo samples. This validates our approach of using real samples obtained for the current task to transfer knowledge backward to previous tasks.
- Tuning one model for all tasks with methods like ER, LAMOL and HAT still faces the issue of forgetting previously learned knowledge (i.e., negative BWT).
- CTR and Comp bring negative interference among tasks, resulting in poor performances. This validates the effectiveness of assigning task-specific parameters for lifelong learning tasks (Adapter and UELL).
- The higher performance of UELL compared to Adapter indicates that our method makes better use of unlabeled data, and the introduced backward augmentation does transfer knowledge from newly arrived tasks to learned tasks.

6.2 Ablation Studies

We conduct ablation studies to verify the effectiveness of each proposed component in UELL. The task orders of STDD and DT are randomly selected. (1) **w/o Unlabel** means no unlabeled data is utilized for each task. Here, we investigate whether unlabeled data bring benefits to supervised lifelong learning. (2) **w/o Selection** skips the confidence selection process and uses all predicted pseudo labeled data to teach the student, i.e., set $\tau = \infty$ in Eq.2; (3) **w/o Interact** ignores the interaction between the teacher and student, i.e., the loss $\mathcal{L}_u^{\text{T-S}}$ is optimized in every training iteration. (4) **w/o Self-Study** removes the self-study loss $\mathcal{L}_u^{\text{S-S}}$ (Eq.3) for the solver optimization; (5) **w/o Back-Aug** means that newly arrived unlabeled data are not utilized

	FT	EWC	MAS	LAMOL	ER	HAT	CTR	Adapter	Comp	**UELL**	MTL
Avg T.P.	223M	125M	125M	125M	223M	73M	75M	1.79M	2.44M	1.79M	223M
Avg-Score	21.34	45.85	48.01	48.21	64.47	55.24	52.07	69.05	48.13	**71.12**	75.67
BWT	-58.24	-13.31	-13.26	-6.611	-4.012	-3.727	-2.310	0.000	-18.42	**0.146**	N/A

Table 2: Results of five SSTD tasks. Each result is an average of five random task orders. UELL is significantly better than other SSLL baselines with p-value ≤ 0.05 under t-test. "Avg T.P." refers to the number of tunable parameters for each task.

	FT	EWC	MAS	LAMOL	ER	Adapter	Comp	**UELL**	MTL
Avg-Score	25.14	30.31	34.88	48.60	50.91	65.99	43.46	**70.75**	75.83
BWT	-52.14	-28.62	-22.13	-2.590	-13.92	0.000	-3.59	**0.393**	N/A

Table 3: Results of five DT tasks. Each result is an average of five random task orders. UELL is significantly better than other SSLL baselines with p-value ≤ 0.05 under t-test.

to promote the learned solvers, i.e., schemes introduced in §4.3 are ignored.

From Table 4, we can see that: (1) Unlabeled data leveraged by UELL greatly improve the supervised lifelong learning. This validates our claim that UELL can effectively overcome the first challenge of SSLL, i.e., exploiting unlabeled data. (2) Removing the pseudo-label selection process makes our student suffer from noisy pseudo labels predicted by the teacher, thus downgrading their lifelong learning performance. This also validates the effectiveness of our confidence-based label selection scheme. (3) Removing interaction between the teacher and student degenerates the performance. It verifies that learning harder and noisier knowledge from unlabeled data is non-trivial. (4) Self-study of the student indeed enhances its capability of utilizing unlabeled data. (5) The backward augmentation encourages the knowledge transfer from new tasks to old tasks, further promoting overall performance. This validates our claim that UELL can effectively tackle the second challenge of SSLL, i.e., leverage unlabeled data to encourage knowledge transfer to previous tasks.

6.3 Data Efficiency Analyses

We analyze the data efficiency of UELL by fixing the unlabeled data and vary the number of labeled data on SSLL. We randomly select a task order from STDD to conduct the analyses. As shown in Table 5, the Avg-Score of UELL increases with the amount of labeled data. The performance of UELL does not significantly degenerate in the few-shot setting. Even with only 50 labeled samples, UELL can still surpass all the baselines that use 100 labeled data in Table 2. Moreover, we also notice that the BWT score of UELL trained with 2000 labeled samples is lower than that of 100 labeled data. We speculate this is because our solver has acquired sufficient knowledge from labeled data, and the knowledge it can gain from unlabeled data of subsequent tasks is limited. In this case, it is hard to perform backward transfer because the learned solvers are already knowledgeable.

Further, we assess the ability of UELL to leverage unlabeled data by varying the amount of unlabeled data while keeping the labeled data fixed. As shown in Table 5, we can see that UELL gets better performances (i.e., higher Avg-Score and BWT) with more unlabeled data. This validates the effectiveness of UELL for leveraging unlabeled data to improve the overall performance of SSLL and encouraging more knowledge transfer from new tasks to previous tasks.

	# Labeled Data			# Unlabeled Data			
	50	100	2000	0	500	2000	10000
Avg-Score	68.64	71.57	75.06	68.13	70.51	71.57	71.92
BWT	0.056	0.144	0.016	0.000	0.059	0.144	0.223

Table 5: The performances of UELL with different number of labeled and unlabeled data.

6.4 Longer Task Sequences Analysis

To verify the ability of UELL to handle more tasks, we combine STDD and DT tasks to form a longer sequence of ten tasks. We compare UELL with its upper bound MTL and three best-performing baselines LAMOL, ER, and Adapter. We randomly select three task orders and report their average performances in Table 6. Our method UELL still outperforms these baselines with a large margin, suggesting that UELL can be generalized to longer

		UELL	w/o Unlabel	w/o Select	w/o Interact	w/o Self-Study	w/o Back-Aug
STDD	Avg-Score	**71.57**	68.13	70.64	68.93	70.97	69.73
	BWT	**0.144**	0.000	0.039	0.089	0.049	0.000
DT	Avg-Score	**73.10**	67.34	71.47	71.82	71.42	72.80
	BWT	**0.713**	0.000	0.189	0.021	0.024	0.000

Table 4: Ablation studies on STDD and DT tasks. "w/o" means removing the corresponding component in UELL.

task sequences.

	LAMOL	ER	Adapter	UELL	MTL
Avg-Score	52.71	61.68	67.52	**69.23**	71.82
BWT	-4.610	-4.295	0.000	**0.235**	N/A

Table 6: Performances of UELL and some baselines under longer task sequence.

6.5 Analyses of Backward Augmentation

For backward augmentation in §4.3, the number of neighbors K when retrieving similar data is important. Hence, we conduct analyses to investigate how the value of K affects UELL's performances. A random task order of STDD is selected to implement the analyses. As shown in Table 7, the overall performance of UELL fluctuates with the value of K. The backward transfer performance generally improves as K increases. However, if K is too large, we are more likely to absorb and retrieve dissimilar data to $\mathcal{D}_k^{\text{Mix}}$ and thus degenerate the model performance.

K	1	3	30	50
Avg-Score	71.35	71.57	71.31	71.28
BWT	0.092	0.144	0.158	0.123

Table 7: Impacts of the number of neighbors K.

6.6 Computation Resource Analysis

We report the number of tunable parameters for UELL and baselines to assess their computation cost (see Tabel 2). UELL utilizes the smallest tunable parameters but achieves the best performances of SSLL.

6.7 Case Study of Pseudo Unlabeled Samples

We present some pseudo unlabeled samples generated by UELL in Appendix B. We can observe that UELL generates high-quality pseudo samples for learned tasks. This benefits from the sufficient data ($\mathcal{D}^s$ and $\mathcal{D}^u$) to learn the language modeling ability for UELL by optimizing $\mathcal{L}^{\text{LM}}$ in Eq. 4.

7 Conclusion

In this paper, we propose a new setting, semi-supervised lifelong language learning (SSLL), where a model learns a sequence of language tasks using both labeled and unlabeled data. We build a novel method UELL to tackle challenges in SSLL. UELL contains a virtual supervision enhanced solver to exploit unlabeled data for each task and a backward augmented learner to encourage knowledge transfer from subsequent tasks to previously learned tasks. Extensive experiments and analyses on language tasks demonstrate the effectiveness of UELL in leveraging unlabeled data, mitigating catastrophic forgetting, and encouraging backward knowledge transfer in the SSLL setting.

Limitations

As our first attempt in the new semi-supervised lifelong language learning (SSLL) setting, our method UELL assumes the unlabeled data of each task are intrinsically related to labeled data. We have not investigated the unlabeled data from general corpus such as Common Crawl [4], BooksCorpus (Zhu et al., 2015) and Wikipedia [5] to improve lifelong language learning with limited labeled data.

Fortunately, pre-training schemes may already provide insights for the above problems. The pretrained T5 checkpoints we use to initialize the UELL model have been pretrained on these general corpus with well-designed losses. We can explore including these pretraining losses in further attempts for the SSLL setting. Pre-training models that are obtained from other corpora (He et al., 2022c; Zhou et al., 2021; Wang et al., 2020; Zheng et al., 2020, 2022; He et al., 2022b; Zheng et al., 2019) may also help to alleviate this issue.

Moreover, our UELL model constructs task-specific adapter modules to prevent forgetting. As a similar approach to adapters, the prompt learning also enables us to share the same backbone

[4]Common Crawl link: https://commoncrawl.org

[5]Wikipedia link: https://huggingface.co/datasets/wikipedia

language model while dynamically allocating task-specific parameters (Liu et al., 2021). Prompts can also be used to implement parameter-efficient lifelong learning schemes. The virtual supervision enhanced solver and backward augmented learner proposed in UELL can be directly combined with prompt learning based approaches. In future works, we aim to explore the prompt-based approach to tackle challenges in SSLL. Some approaches on few-shot learning (Zhao et al., 2022a) can also be applied in our SSLL setting.

Ethics Statement

This work does not present any direct ethical issues. In our paper, the method UELL is proposed to cope with a more realistic setting, semi-supervised lifelong language learning, where the model learns sequentially arrived language tasks that are partially labeled. All our experiments are conducted on publicly available datasets. All terms for using these datasets are strictly followed in our study. The metrics used in our paper are automatic and do not involve manual labor.

Acknowledgement

Research on this paper was supported by Alibaba Group through Alibaba Research Intern Program and Hong Kong Research Grants Council (Grant No. 16204920).

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *stat*, 1050:21.

Dhanajit Brahma, Vinay Kumar Verma, and Piyush Rai. 2021. Hypernetworks for continual semi-supervised learning. *arXiv preprint arXiv:2110.01856*.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135.

Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2020. Lifelong language knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2914–2924.

Yi Dai, Hao Lang, Yinhe Zheng, Fei Huang, Luo Si, and Yongbin Li. 2022. Lifelong learning for question answering with hierarchical prompts. *arXiv preprint arXiv:2208.14602*.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. 2019. Learning without memorizing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5138–5146.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. 2017. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022a. Space-2: Tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 553–569.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022b. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022c. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul A Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

Fei Mi, Liangwei Chen, Mengjie Zhao, Minlie Huang, and Boi Faltings. 2020. Continual learning for natural language generation in task-oriented dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3461–3474.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020b. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152.

Chengwei Qin and Shafiq Joty. 2022. Lfpt5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. *ICLR 2022*.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P Lillicrap, and Greg Wayne. 2019. Experience replay for continual learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 350–360.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pages 4528–4537. PMLR.

Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*.

James Smith, Jonathan Balloch, Yen-Chang Hsu, and Zsolt Kira. 2021. Memory-efficient semi-supervised continual learning: The world is its own replay buffer. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning. In *International Conference on Learning Representations*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Liyuan Wang, Kuo Yang, Chongxuan Li, Lanqing Hong, Zhenguo Li, and Jun Zhu. 2021. Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5383–5392.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

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)*, pages 6382–6388.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020a. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020b. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. 2019. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485.

Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2022. Continual sequence generation with adaptive compositional modules. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3653–3667.

Yingxiu Zhao, Zhiliang Tian, Huaxiu Yao, Yinhe Zheng, Dongkyu Lee, Yiping Song, Jian Sun, and Nevin Zhang. 2022a. Improving meta-learning for low-resource text classification and generation via memory imitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–595, Dublin, Ireland. Association for Computational Linguistics.

Yingxiu Zhao, Yinhe Zheng, Zhiliang Tian, Chang Gao, Bowen Yu, Haiyang Yu, Yongbin Li, Jian Sun, and Nevin L. Zhang. 2022b. Prompt conditioned vae: Enhancing generative replay for lifelong learning in task-oriented dialogue. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. MMChat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5778–5786, Marseille, France. European Language Resources Association.

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.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, et al. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.

Zhi-Hua Zhou and Ming Li. 2010. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Experiments Details

A.1 Datasets and Metrics

Details of the datasets we use in our studies are listed below. Five datasets (tasks) from decaNLP (McCann et al., 2018; Sun et al., 2019):

- Question Answering – Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016): It consists of context, questions, and answers. The context is paragraphs from English Wikipedia, and the answers are spanned from its corresponding question paragraphs.

- Semantic Parsing – WikiSQL (Zhong et al., 2017): WikiSQL provides logical forms along with natural language utterances.

- Sentiment Analysis – Stanford Sentiment Treebank (SST) (Radford et al., 2017): It consists of movie reviews with its answers, including positive and negative binary options.

- Semantic Role Labeling – QA-SRL (He et al., 2017): It is a question answering form of the SRL task.

- Goal-Oriented Dialogue – English Wizard of Oz (WOZ) (Wen et al., 2016): WOZ is a restaurant reservation task that provides a predefined ontology of a series of information for helping an agent to make reservations for customers.

Five text classification datasets (tasks) from MBPA++ (d'Autume et al., 2019; Sun et al., 2019):

- AGNews: News articles to be classified into 4 classes.

- Yelp and Amazon: Customer reviews and ratings on Yelp and Amazon. Both datasets include 5 classes.

- DBPedia: Articles and their corresponding categories on Wikipedia, including 14 classes.

- Yahoo: Questions and answers on the Yahoo! platform, including 10 classes.

A.2 Details of Baselines Implementation

For LAMOL, we also use task-specific generation tokens to perform the generative replay. The sampling ratio of pseudo samples is set to 0.2. For ER, we store real samples of labeled data and unlabeled data with a ratio of 0.1 for each task. To adapt compared baselines to SSLL, models trained on labeled data are used to generate pseudo labels of unlabeled data and are further trained with the mixture of labeled data and pseudo labeled data like (Xie et al., 2020b).

B Case Study of Pseudo Unlabeled Data

We show a few pseudo unlabeled data of three tasks generated by UELL in Table 8.

Sentiment Analysis
1. The performance is priceless. 2. This humbling film, in all its minutiae and in its ambiguity, is simply a waste of money. 3. This is one of those films that sneaks up on you and stays with you long after you' ve left the theatre.
News Classification
1. Oil Spite Changes (News) Petroleum companies are changing their share price target to become cheaper, a new research study said Wednesday. 2. In a bid to reach the future of health care, the government recently endorsed the use of the halo method to treat chronic diseases. 3. At least two teams make headlines to help a new scottish league team to win their first title defending champion Al Hamdat this December, as a defensive midfielder.
Goal-Oriented Dialogue
1. I would like a cheap restaurant in the north part of town. 2. Can you tell me the phone number of the Chinese restaurant? 3. Yes, I need the address and phone number please.

Table 8: Generated pseudo unlabeled data for three tasks.