Ask Question First for Enhancing Lifelong Language Learning

Han Wang[1,2] and **Ruiliu Fu**[1,2] and **Xuejun Zhang**[1,2] and **Jun Zhou**[1,2] and **Qingwei Zhao**[1,2]

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China.

[2]University of Chinese Academy of Sciences, Beijing, China

{wanghan, furuiliu, zhangxuejun, zhoujun, qzhao}@hccl.ioa.ac.cn

Abstract

Lifelong language learning aims to stream learning NLP tasks while retaining knowledge of previous tasks. Previous works based on the language model and following data-free constraint approaches have explored formatting all data as "begin token (B) + context (C) + question (Q) + answer (A)" for different tasks. However, they still suffer from catastrophic forgetting and are exacerbated when the previous task's pseudo data is insufficient for the following reasons: (1) The model has difficulty generating task-corresponding pseudo data, and (2) A is prone to error when A and C are separated by Q because the information of the C is diminished before generating A. Therefore, we propose the Ask Question First and Replay Question (AQF-RQ), including a novel data format "$BQCA$" and a new training task to train pseudo questions of previous tasks. Experimental results demonstrate that AQF-RQ makes it easier for the model to generate more pseudo data that match corresponding tasks, and is more robust to both sufficient and insufficient pseudo-data when the task boundary is both clear and unclear. AQF-RQ can achieve only 0.36% lower performance than multi-task learning.

1 Introduction

Lifelong learning is the capacity of human beings to acquire, reconstruct, strengthen, and transfer knowledge (Ring, 1997). Human beings can learn new knowledge while consolidating old knowledge by first detecting and learning the distinctions between old and new knowledge and then simplifying the old and new knowledge based on leveraging the common points. This concept is critical for encouraging machines to learn NLP tasks in a similar way that people do. In the application of NLP, new data are continuously acquired and categorized as either new data for existing tasks or new data for new tasks. For new data of existing tasks, the traditional method, known as isolated learning (Chen
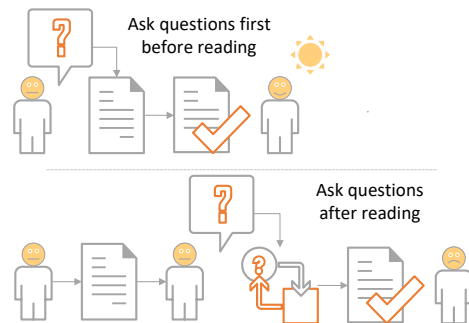

Figure 1: Illustration of the difference between asking questions before reading and asking after reading.

and Liu, 2018), is to retrain the model with the old data appended with new data. For data from new tasks, multi-task learning (MTL) integrates data from previous and new tasks to retrain the model. Both of these methods are limited to assuming that all new and old data can be obtained during training. However, the reality is that tasks are acquired and trained in the stream, which makes the model suffer from catastrophic forgetting (Ring, 1997; McCloskey and Cohen, 1989; French, 1999) (i.e. forgetting previously learned tasks/knowledge).

Lifelong language learning (LLL), which this paper focuses on, aims to learn a stream of NLP tasks with lifelong learning. LAMOL (Sun et al., 2019) has recently proposed implementing a language model for LLL by formatting all data as QA-style and generating pseudo data instead of real data for previous tasks in order to prevent catastrophic forgetting. Many works (Chuang et al., 2020; Sun et al., 2020; Kanwatchara et al., 2021) have investigated how to improve it through methods that require more computing resources or additional parameters (e.g. knowledge distillation or adding sub-networks). In this paper, we refer to these works as LAMOL-based methods. In LAMOL-based methods, each example is consist of four segments: task-specific or task-independent token

4610

[TASK]/[GEN] (B), context (C), question (Q), and answer (A). Then each example is formatted as "B+C+Q+A" ("$BCQA$"). LAMOL has a gap between MTL that is regarded as the upper bound of LLL. Other LAMOL-based methods require more computing resources or additional parameters but still have a gap between MTL especially pseudo samples is unsufficient. Therefore, the "$BCQA$" format is useful but not train-efficient enough for reasons below: (1) generating task-corresponding pesudo data is hard because the similar C can be found in distinct tasks; (2) the C's information is diminished when A is generated due to the fact that A and C are separated by Q.

To generate more task-corresponding pseudo data for previous tasks and tighten the relationship between the C and the A, we proposed the Ask Question First (AQF) and Replay Question (RQ).

The AQF formats all data into a novel format: "B+Q+C+A" ("$BQCA$"), which is consistent with human reading comprehension behavior. Questions are more than just questions; they direct our learning. As is shown in Fig.1, when people do reading comprehension, they usually read the questions first, then read the articles with the questions, and pay attention to what can answer the questions while reading. This is an efficient reading and learning method, which known as metacognition (Flavell, 1979), has been researched in the field of education and psychology. There are two benefits to applying the "$BQCA$" format: (1) It's easier to generate pseudo-data that matches the corresponding task because the Q has different but limited types for each task. (2) The model is more stable when there isn't enough pseudo data because the A is right next to the C. Furthermore, the model can pay more attention to important information in the C with the help of the Q.

The RQ introduce a novel training task to help model generate more task-corresponding pseudo data. In the "$BQCA$" format, generating the correct Q is crucial for generating task-corresponding pseudo-data. However, the Q of the previous may be covered by the new task since we cannot predict the number and type of Q of the new task. In order to strengthen the generation of the Q of the previous task, we generate pseudo-problems of the previous task to train the model.

The contributions of our paper are listed below: (1) We proposed the Ask Question First and Re-

play Question (AQF-RQ[1]) to alleviate catastrophic forgetting when the pseudo data is sufficient and insufficient without additional computation resources and parameters. (2) We proposed a novel data format "$BQCA$" to make data train-efficient and generating corresponding pseudo data easier. (3) We proposed a novel training task to help the model generate correct questions and then generate more task-corresponding pseudo data.

2 Related Work

Lifelong language learning (LLL) is an essential step in promoting the realization of general artificial intelligence in the field of NLP. (Liu et al., 2019; Mi et al., 2020; Huang et al., 2021) have studied LLL on a single type of NLP task by regularization or replaying real data. Recently, LAMOL (Sun et al., 2019) uses a language model (LM) to learn various kinds of NLP tasks in QA-style. In LAMOL, the pseudo-data generated by the model is trained together with the new task to alleviate catastrophic forgetting. Many works explored the enhancement of LAMOL with additional computation resources or adding sub-networks. L2KD (Chuang et al., 2020), DnR (Sun et al., 2020) and DFSD (Wang et al., 2022) distilled parts or all layers of the model to improve LAMOL. ARPER (Mi et al., 2020) applied regualarization on parameters with prioritized exemplar replay. Rational-LAMOL (Kanwatchara et al., 2021) applied critical freezing guided by rationale information which is obtained by human or unsupervised rationale generation (Chang et al., 2020). (Madotto et al., 2021) applied Adapter (Houlsby et al., 2019) to plug into pretrain language model. Those are effective but require more computational resources (e.g., distillation and regularization) or new parameters (such as adding sub-networks for new tasks). We aspire to improve without increasing our resources. Consequently, based on the metacognition (Flavell, 1979; Bowler, 2010; Braithwaite and Sprague, 2021) proposed in education and psychology, we propose a more training-efficient and robust data format: first ask questions, then observe the context, finally answer the questions. In addition, we introduce replay questions to strengthen the model's attention to questions.
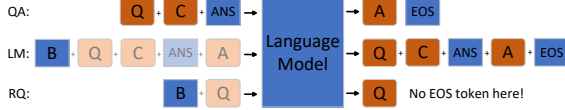
---

[1] https://github.com/CodeHan/AQF-RQ

Figure 2: Illustration of QA/LM/RQ task with the "*BQCA*" format.

3 Methodology

In this section, we first introduce LAMOL in Section 3.1. Then, our proposed Ask Question First (AQF) is detailed in Section 3.2. Thirdly, we introduce the Replay Questions in Section 3.3. Finally, we summarize the training objectives in our paper in Section 3.4.

3.1 LAMOL

LAMOL (Sun et al., 2019) proposed using a single language model for lifelong language learning. A stream of various NLP tasks is learned by GPT-2 (Radford et al., 2019) with joint training of the language model (LM) task and the question-answering (QA) task. In LAMOL, all tasks are formatted QA-style. Each example can consist of four segments: task-specific or task-independent token [TASK]/[GEN] (B), context (C), question (Q), and answer (A). In this way, the LM task is to generate "$C+Q+A$" by inputting [TASK]/[GEN], and the QA task is to generate "A" by inputting "$C+Q$". The loss of LAMOL is calculated as below:

$$\mathcal{L}_{LAMOL} = \mathcal{L}_{QA} + \lambda \mathcal{L}_{LM} \qquad (1)$$

where λ is the weight of the LM task.

With the help of the LM task, the [TASK]/[GEN] can be used as the first token to input the model and generate the pseudo sample by greedy decoding. From the second task on, pseudo samples of previous tasks are generated in this way and jointly trained with the new task dataset to alleviate catastrophic forgetting. Let denote the set of pseudo samples as $\mathcal{P}_i = \{ps_j | i, j \in N_+, 1 \leq j < i\}$, where ps_j represents the pseudo data of j-th task.

The number of pseudo samples of previous tasks is determined by the size of the new task and a hyperparameter sampling ratio $\gamma \in (0, 1]$. Assuming that $T = \{T_1, T_2, \ldots, T_N\}$ represents N tasks to be learned. Let $|D_t|, t \in [2, N]$ denote the size of t-th task. LAMOL generates $|ps_i| = \frac{\gamma}{t-1}|D_t|$ pseudo-samples for each previously learned task. Then, D_t is joined with $\mathcal{P}_t$ to train the model. However, the performance of LAMOL still has a gap between

Case ID	Content	Cor.
1	$B^{T_1}+C^{T_1}+Q^{T_1}+A^{T_1}$	✓
2	$B^{T_1}+C^{T_2}+Q^{T_2}+A^{T_2}$	✗
3	$B^{T_2}+C^{T_2}+Q^{T_2}+A^{T_2}$	✓
4	$B^{T_2}+C^{T_1}+Q^{T_1}+A^{T_1}$	✗

Table 1: The cases of the pseudo sample after the model learning two tasks in turn with "*BCQA*" format. Cor. is the abbreviation of "corresponding".

multi-task learning (MTL), which is regarded as the upper bound of lifelong learning, since not each pseudo sample is tied to [TASK]/[GEN]. In this paper, we propose Ask Question First (AQF) and replaying questions to make data train-efficient and shorten the gap between MTL.

3.2 Ask Question First

Ask Question First (AQF) is a novel data format that makes the model imitate the process of human reading comprehension to learn knowledge from the training set efficiently. LAMOL proposed a data format, "$B+C+Q+A$" ("*BCQA*"), which is useful but not train-efficient enough. All pseudo samples are generated starting with [TASK]/[GEN]. However, [TASK]/[GEN] is not strongly tied with the C of the tasks because the context is complex and volatile for the task. Different tasks can have similar contexts. Therefore, [TASK]/[GEN] is easy to be biased to the new task resulting in pseudo samples not corresponding to their tasks when more tasks are learned. These non-corresponding pseudo-samples jointly trained with the new task will aggravate the catastrophic forgetting of the model. Otherwise, the C's information is diminished when A is generated due to the fact that A and C are separated by Q.

To generate more corresponding pseudo samples, we propose AQF format all data as "$B+Q+C+A$" ("*BQCA*") which is simple yet efficient to make the model learn knowledge from the training set efficiently. We choose the task-specific token to analyze why our proposed "*BQCA*" is better than the naive "*BCQA*". Assuming that the model has learned two tasks T_1 and T_2, pseudo samples of T_1 and T_2 needed to be generated before training the new task. Two types of task-specific tokens B_1 and B_2 are inputted to the model to generate pseudo samples with greedy decoding. There are four main kinds of pseudo samples that can be obtained as shown in Table 1. Case 1 and Case 3

are nice pseudo samples that can alleviate catastrophic forgetting, but Case 2 and Case 4 are terrible ones that can aggravate catastrophic forgetting. Ideally, pseudo samples, which we denote as task-corresponding pseudo-samples $\mathcal{P}_c$, like Case 1 and Case 3 are what we expect to generate. However, in fact, due to catastrophic forgetting, Case 2 and Case 4 also appear. These pseudo samples like case 2 and case 4 are denoted as the not task-corresponding samples $\mathcal{P}_{nc}$. The probability of generating $\mathcal{P}_{nc}$ is positively correlated with the severity of catastrophic forgetting. Therefore, the model will form a vicious circle: forgetting old tasks and generating $\mathcal{P}_{nc}$ with greater probability. For pseudo samples start with task-specific token, the ps_i consists of corresponding pseudo data ps_i^c and not corresponding pseudo data ps_i^{nc} (i.e. $ps_i = \{ps_i^c, ps_i^{nc}\}$).

In LAMOL-based methods, they used GPT-2 (Radford et al., 2019) for experiments. GPT-2 is an autoregressive model that can only obtain information before the current position, and then predict the words. Therefore, the probability of our proposed "*BQCA*" format and naive "*BCQA*" format can be calculated by Eq.(2) and Eq.(3), respectively:

$$P_{AQF} = P(BQCA) \qquad (2)$$
$$= P(A|BQC)P(C|BQ)P(Q|B)P(B)$$
$$P_{naive} = P(BCQA) \qquad (3)$$
$$= P(A|BCQ)P(Q|BC)P(C|B)P(B)$$

As shown in Eq.(2) and Eq.(3), $P(B)$ is a constant because it is given as the first token. Next, $P(Q|B)$ and $P(C|B)$ affect the quality of the generated pseudo data. To obtain $\mathcal{P}_c$, the Q or C should be generated to be corresponding to the given B, then the remaining generation has a greater probability correspond to B.

Then we first consider $P(Q|B)$ and $P(C|B)$. Let Θ_N denote the parameters of the model after learning N tasks in turn. Assuming two tasks have been learned on the condition of sufficient pseudo samples (i.e. large γ) in turn, $P(Q|B)$ and $P(C|B)$ can be written as, $P(Q|B, \Theta_2)$ and $P(C|B, \Theta_2)$. In the dataset of a task, there are thousands of kinds of C, but there are only limited kinds of Q. For a simplest classification task, only one fixed Q is needed. For example, in the SST (Radford et al., 2017), all Q is "*is review is positive or negative?*". For complex tasks, such as SRL (He et al., 2017), it is mainly based on limited types of Q such as *what*, *where*, *why*, *when*,

how etc, but it is still smaller than that of C. Therefore, $P(Q|B, \Theta_2)$ and $P(C|B, \Theta_2)$ are inversely proportional to the vocabulary sizes of Q and C, respectively. Let $\mathcal{V}_Q = \{w_i^Q | i \in [1, |\mathcal{V}_Q|]\}$ and $\mathcal{V}_C = \{w_i^C | i \in [1, |\mathcal{V}_C|]\}$ denote the vocabulary of Q and C, respectively. w_i^Q and w_i^C denotes the word in the vocabulary of Q and C, respectively. According to the above analysis, it can be obtained that $|\mathcal{V}_Q| \ll |\mathcal{V}_C|$. For most words, since their probability is inversely proportional to the size of the vocabulary, we can conclude that $P(w_Q) \gg P(w_C)$. Slight noise perturbations, such as shifts from the data distribution of the new task, can make $P(w_C)$ more susceptible than $P(w_Q)$. On the other hand, different tasks may have similar C due to a large number of similar common phrases, but the Q is almost different in what the questions pay attention to. Therefore, we can conclude that:

$$P_{B_1}(w_Q^{T_1}) \geq P_{B_1}(w_Q^{T_2}) > P_{B_1}(w_C^{T_1}) \geq P_{B_1}(w_C^{T_2}) \tag{4}$$

$$P_{B_2}(w_Q^{T_2}) \geq P_{B_2}(w_Q^{T_1}) > P_{B_2}(w_C^{T_2}) \geq P_{B_2}(w_C^{T_1}) \tag{5}$$

where the subscript B_1 or B_2 means the probability on the condition of B_1 or B_2. Then, we can conclude that:

$$P(Q^{T_1}|B^{T_1}, \Theta_2) > P(C^{T_1}|B^{T_1}, \Theta_2) \tag{6}$$
$$P(Q^{T_2}|B^{T_2}, \Theta_2) > P(C^{T_2}|B^{T_2}, \Theta_2) \tag{7}$$

It can be concluded that generating Q first is more stable than generating C.

On the condition that the Q or C is corresponding to the B, we analysis the $P(C|BQ, \Theta_2)$ and $P(Q|BC, \Theta_2)$. With the help of GPT-2, a Transformer decoder-based model, the current token can pay more attention to the information before the current position. When the previous adjacent content contains enough information, the generated token strongly correlates with the previous content. Therefore, $P(C^{T_1}|B^{T_1}Q^{T_1}, \Theta_2)$ is close to $P(Q^{T_1}|B^{T_1}C^{T_1}, \Theta_2)$, and $P(C^{T_2}|B^{T_2}Q^{T_2}, \Theta_2)$ is close to $P(Q^{T_2}|B^{T_2}C^{T_2}, \Theta_2)$.

Finally, $P(A|BQC)$ and the $P(A|BCQ)$ are trained with the QA task. It can be ignored from the perspective of generating task-corresponding pseudo-samples. However, correct "*A*" is easier to obtain on the condition of correct "*BQC*" than correct "*BCQ*" because the "*C*" is right next to "*A*" in "*BQC*". In "*BQC*", "*Q*" can help the model pay more attention to the important information in "*C*"

4613

and then have a larger probability of generating the correct "A".

From the above analysis, we can conclude that $P(Q|B)$ and $P(C|B)$ play main role on generating task-corresponding pseudo-samples $\mathcal{P}_c$ according to P_{AQF} and P_{naive}, respectively. As shown in Eq.(6) and Eq.(7), $P(Q|B)$ are better than $P(C|B)$ for generating $\mathcal{P}_c$ after sufficient pseudo samples being joint trained.

3.3 Replay Questions

Replay Questions (RQ) is proposed to strengthen the probability of generating task-corresponding Q when given B. Based on Eq.(4), $P_{B_1}(w_Q^{T_1})$ greater than or equal to $P_{B_1}(w_Q^{T_2})$ on the condition of sufficient pseudo samples being joint trained. However, the situation $P_{B_1}(w_Q^{T_2}) > P_{B_1}(w_Q^{T_1})$ may happen when the amount of pseudo samples is much less than that of the new task (i.e. $|D_2| \gg \gamma|D_2|$). Since $P(Q|B)$ plays a main role in P_{AQF}, we propose to replay questions to make $P(Q_i|B_i)$ larger than $P(Q_{j \neq i}|B_i)$.

As shown in Fig.2, RQ is to generate the questions of pseudo samples that start with [TASK]/[GEN] but have no end token. Since the object of RQ is to make the model generate task-corresponding pseudo-samples, RQ should not supply the information about the end of the sentence.

3.4 Training

In summary, our proposed AQF is applied to both the LM task and the QA task. Otherwise, we expand the training objective with our proposed RQ. The final training objective is shown as Eq.(8).

$$\mathcal{L} = \mathcal{L}_{QA}^{AQF}(D, \mathcal{P}) + \lambda \mathcal{L}_{LM}^{AQF}(D, \mathcal{P}) + \eta \mathcal{L}_{RQ}^{AQF}(\mathcal{P})$$
(8)

where λ and η denotes the weight of the LM task and the RQ task.

4 Experiment Setup

4.1 Datasets

To be comparable with previous works, we chose four different kinds of tasks from DecaNLP (Mc-Cann et al., 2018) and five sequence-generation tasks (Chuang et al., 2020) from different domains. Details are summarized in the Table 7 in Appendix A. The SQuAD is a question-answering dataset with 12 main types of questions and other countless types of questions. SST is a sentiment analysis dataset with one question. SRL is a semantic role labeling dataset with 6 types of questions.

Tasks \ Format	BCQA	BQCA (ours)
SQuAD	72.3	**81.1**
SST	90.9	**92.2**
SRL	70.4	**73.9**
WOZ	84.9	**86.7**
E2ENLG	48.8	**49.2**
RNNLG (rest)	64.0	**64.6**
RNNLG (hotel)	65.4	**66.4**
RNNLG (tv)	70.8	**71.6**
RNNLG (laptop)	73.0	**73.2**

Table 2: The results on each task with single task learning (only the QA task is trained). The format "$BCQA$" was applied in previous LAMOL-based methods. Better performance in boldface.

WOZ is a goal-oriented dialogue dataset with one question. E2ENLG (Novikova et al., 2017) and RNNLG (rest/hotel/tv/laptop) (Wen et al., 2015)are sequence generation tasks with one question for different domains.

4.2 Baselines

(i) **Finetune**: Finetune GPT-2 individually based on the task order. (ii) **LAMOL** (Sun et al., 2019): Training the model with the "$BCQA$" format data. $LAMOL_G$ and $LAMOL_T$ indicate that the begin token is the task-independent token [GEN] and the ask-specified token [TASK], respectively. $LAMOL_R$ indicates that the model replays real data from previous tasks instead of pseudo data. (iii) **L2KD** (Chuang et al., 2020): An improved version of LAMOL with distillation. Firstly, training a single-task model on the new task to obtain the teacher model. Then the teacher model is distilled to the model trained on the previous tasks. This model uses the task-specified token. All the LAMOL-based baselines were studied upon the "$BCQA$" format of LAMOL. We only select L2KD on behalf of other LAMOL-based baselines because of a lack of computation resources. (iv) **Multitask**: Training all tasks simultaneously. Multitask learning is often regarded as the upper bound of lifelong learning.

4.3 Implementation Detail

For fairness comparation. we implement experiments following LAMOL[2]. GPT-2 with 12 layers is selected as the language model. All experiments are run on a single Tesla P100 (12GB). Each task is trained for 9 epochs. The pseudo data is generated

---

[2]https://github.com/jojotenya/LAMOL

by greedy deocde. The other hyperparameter settings are the same as for LAMOL and are detailed in the Table 6 in the Appendix B.

5 Experiments

In this section, we experiment our proposed Ask Question First and Replay Question (AQF-RQ) an on three settings: (1) single task learning in Section 5.1; (2) different types of tasks in Section 5.2; (3) the same types of tasks in different domains in Section 5.3. Finally, we explore the effectiveness of our proposed AQF in Section 5.4.

5.1 Single Task

To validate our proposed AQF that format all task into new data format "$BQCA$", we experiment on each dataset independently. Each task is only trained by the QA task. As shown in Table 2, our proposed "$BQCA$" format beats "$BCQA$" format which is applied in previous LAMOL-based methods on each task. It exhibits that our "$BQCA$" format has the ability to increase the performance of a variety of different kinds of tasks, particularly those that were originally question-answering types (e.g. SQuAD, SRL). The SQuAD is improved significantly with the "$BQCA$" format scoring 8.8 percentage points higher than the "$BCQA$" format. For SRL, 3.5 percentage points of improvement come from the "$BQCA$" format compared with the "$BCQA$" format. For SST and WOZ, the "$BQCA$" format can improve by 1.3 and 1.8 percentage points respectively. For five generation tasks, "$BQCA$" format is slightly better than "$BCQA$".

There are two reasons why our proposed "$BQCA$" format can improve the aforementioned tasks to a certain extent: (1) Due to the characteristics of autoregression and the mask attention mechanism of the Transformers Decoder, first Q and then C can enable the model to obtain more accurate attention information from C based on Q. (2) The A is right next to the C, which reduces the information loss caused by the excessive length of the historical text in comparison to "$BCQA$" where the A and the C are gapped by the Q.

5.2 Different Types of Tasks

For a fair comparison, we conduct experiments on three different tasks in DecaNLP following LAMOL (Sun et al., 2019). To observe the performance in the case of sufficient and insufficient pseudo data, we select the sampling ratio $\gamma = 0.2$

and $\gamma = 0.05$ for experiments. Meanwhile, we conduct experiments on the condition of task-specific token [TASK] (denoted by the subscript T) and task-independent token [GEN] (denoted by the subscript G) to verify if our proposed AQF-RQ is robust when the task boundary is clear or unclear.

5.2.1 Performance

As shown in Table 3, Finetune, which is a baseline for other methods, suffers serious catastrophic forgetting and has large gap bewteen multitask learning. Let's first observe the situation where the pseudo data is sufficient (i.e. $\gamma = 0.2$). $LAMOL_G^{0.2}$ and $LAMOL_T^{0.2}$ have good performance and are similar, indicating that in this case, the "$BCQA$" format is robust to clear and unclear task boundaries, but still 3 percentage points lower than MTL. However, AQF-RQ does better than LAMOL by 2.7 percentage points when the boundary is clear and by 2.3 percentage points when the boundary is not clear. AQF-RQ is only 0.3-0.6% worse than MTL. It is worth noting that AQF-RQ outperforms multi-task learning in the three task orders: SST-SRL-WOZ, SST-WOZ-SRL, and WOZ-SRL-SST. This indicates that AQF-RQ can not only better alleviate catastrophic forgetting but also further strengthen forward transfer between tasks.

When the pseudo data is insufficient (i.e. $\gamma = 0.05$), the performance of LAMOL drops significantly, falling 6.5-7.6 percentage points lower than that of multitask learning. Furthermore, $LAMOL_T^{0.05}$ differs from $LAMOL_G^{0.05}$, indicating that LAMOL is not robust enough for clear and unclear task boundaries when the amount of pseudo data is insufficient. Although LAMOL's job boundary is apparent, our proposed AQF-RQ can improve by 4.7 and 5.5 percentage points compared to LAMOL. This illustrates that even when the amount of pseudo data is insufficient, AQF-RQ is still robust to both clear and unclear task boundaries, as evidenced by $AQF\text{-}RQ_T^{0.05}$ and $AQF\text{-}RQ_G^{0.05}$. It is also worth noting that $AQF\text{-}RQ_T^{0.05} >$ $AQF\text{-}RQ_G^{0.05} > LAMOL_G^{0.2} > LAMOL_T^{0.2}$, which means that AQF-RQ is more data-efficient and still performs better than LAMOL with a 75% reduction in the amount of pseudo data.

5.2.2 Distribution of Pseudo Data

The another main objective of our proposed AQF-RQ is to generate more corresponding-task pseudo data. Ideally, the ratio of pseudo data for each task is 1:1, which means $|ps_i| = |ps_j|, i \neq j$. We can

Methods	SST SRL WOZ	SST WOZ SRL	SRL SST WOZ	SRL WOZ SST	WOZ SST SRL	WOZ SRL SST	Avg.	Std.
Finetune	42.8	25.2	58.8	32.2	25.6	36.2	36.8	11.6
$\text{LAMOL}_T^{0.05}$	77.3	76.9	78.1	74.7	73.4	75.8	76.0	1.6
$\text{LAMOL}_G^{0.05}$	79.6	78.9	73.1	73.7	68.6	75.7	74.9	3.4
$\text{AQF-RQ}_T^{0.05}$	81.6	80.8	<u>80.3</u>	<u>80.0</u>	<u>80.6</u>	80.7	<u>80.7</u>	<u>1.1</u>
$\text{AQF-RQ}_G^{0.05}$	<u>82.2</u>	<u>81.3</u>	78.9	79.0	79.2	<u>81.7</u>	80.4	2.9
$\text{LAMOL}_T^{0.2}$	79.4	79.9	80.1	78.7	79.8	79.0	79.5	0.5
$\text{LAMOL}_G^{0.2}$	80.0	80.7	79.6	78.7	78.4	80.5	79.7	0.8
$\text{AQF-RQ}_T^{0.2}$	82.2	**83.0**	**81.4**	**81.6**	82.2	**82.9**	**82.2**	**0.5**
$\text{AQF-RQ}_G^{0.2}$	**82.8**	82.7	80.9	80.9	**82.3**	82.4	82.0	0.8
Multitask	82.5							

Table 3: The results on [SST, SRL, and WOZ]. Each column is the average score of a task order on three tasks. Avg. means that the average score on 6 task orders of [SST, SRL, and WOZ]. In the column "Methods", The subscripts T and G indicate the task-specific token [TASK] and task-independent token [GEN], respectively. The superscript indicates the value of sampling ratio γ.

Methods	SST SRL WOZ	SST WOZ SRL	SRL SST WOZ	SRL WOZ SST	WOZ SST SRL	WOZ SRL SST
	$|ps_1| : |ps_2|$					
$\text{LAMOL}_G^{0.05}$	27:99	23:297	7:119	3:343	19:301	6:340
$\text{AQF-RQ}_G^{0.05}$	17:109	12:308	22:104	29:317	70:250	168:178
$\text{LAMOL}_G^{0.2}$	292:216	319:963	128:380	284:1100	225:1057	388:996
$\text{AQF-RQ}_G^{0.2}$	253:255	255:1027	239:269	603:781	712:570	648:736

Methods	SST SRL WOZ	SST WOZ SRL	SRL SST WOZ	SRL WOZ SST	WOZ SST SRL	WOZ SRL SST
	$|ps_1^c| : |ps_1^{nc}|/|ps_2^{nc}| : |ps_2^c|$					
$\text{LAMOL}_T^{0.05}$	54:9/1:62	11:149/2:158	51:12/0:63	22:151/1:172	60:100/0:160	139:34/0:173
$\text{AQF-RQ}_T^{0.05}$	63:0/0:63	160:0/0:160	63:0/2:61	173:0/0:173	160:0/0:160	172:1/1:172
$\text{LAMOL}_T^{0.2}$	247:7/9:245	478:163/5:636	250:4/0:254	537:155/6:686	635:6/4:637	678:14/5:687
$\text{AQF-RQ}_T^{0.2}$	254:0/0:254	641:0/0:641	254:0/0:254	692:0/0:692	641:0/0:641	692:0/0:692

Table 4: The results of pseudo data distribution after learning two tasks on [SST, SRL, and WOZ]. ps_i represents the pseudo data of i-th task. The subscripts c and nc indicate whether the pseudo data is correspond to the task or not.

determine which task the pseudo data belongs to according to the Q because the Q for various tasks is distinct (see Table 7). For task-specific token [TASK], a pseudo sample is task-corresponding if the Q is correspond to the [TASK]. For task-indepedent token [GEN], we can not judge whether a pseudo sample corresponds to the task because of unclear task boundary. In this paper, we applied this Q-based judgment method to count the distribution of pseudo data after learning the second task.

As shown in Table 4, AQF-RQ can generate more pseudo data of the first task for most task orders when applying task-indepedent token [GEN]. Looking at the Table 3 together, it can be observed that the performance will be better if the ratio of the pseudo data of different tasks is closer to 1:1. For SRL-SST-WOZ, SRL-WOZ-SST, WOZ-SST-SRL, and WOZ-SRL-SST, AQF-RQ generates more pseudo data than LAMOL for the first

task; hence, AQF-RQ outperforms LAMOL significantly. When $\gamma = 0.05$, AQF-RQ is 5.3-10.6 percentage points higher than LAMOL. When $\gamma = 0.05$, AQF-RQ is 1.3-3.9 percentage points higher than LAMOL. This is due to the fact that, under the AQF-proposed "$BQCA$" format, Q can be used to help the model pay more attention to the useful details in C, and then A being immediately next to C can further generate A more accurately.

The benefit of AQF-RQ to produce pseudo data is also significant when the sentence starts with the task-specific token [TASK]. As shown in the Table 4, the pseudo samples generated by AQF-RQ are task-corresponding in the majority of task orders, except SRL-SST-WOZ and WOZ-SRL-SST, where one or two pseudo samples are not task-corresponding. The amount of not task-corresponding pseudo samples in AQF-RQ is substantially smaller than in LAMOL, giving AQF-RQ a 1.3-3.9 percentage point advantage over LAMOL.

Methods	SST SRL WOZ	SST WOZ SRL	SRL SST WOZ	SRL WOZ SST	WOZ SST SRL	WOZ SRL SST	Avg.
LAMOL$_T$	79.4	79.9	80.1	78.7	79.8	79.0	79.5
w/ AQF	**82.1**	**83.0**	**81.3**	**80.9**	**80.7**	**81.2**	**81.5**
LAMOL$_G$	80.0	80.7	79.6	78.7	78.4	80.5	79.7
w/ AQF	**82.3**	**81.5**	**81.3**	**80.4**	**78.8**	**81.3**	**80.9**
LAMOL$_R$	81.8	80.6	81.6	81.2	80.4	80.5	81.0
w/ AQF	**82.5**	**81.9**	**82.3**	**82.4**	**82.7**	**82.7**	**82.4**
L2KD	80.1	79.6	79.5	79.7	79.9	**80.4**	79.9
w/ AQF	**81.9**	**82.4**	**80.3**	**80.8**	**80.1**	80.1	**80.9**
Multitask*				81.5			
w/ AQF				**82.5**			

Table 5: The results of effectiveness of the Ask Question First (AQF) on [SST, SRL, and WOZ] when sampling ratio $\gamma = 0.2$. Multitask* represents that model applies the "*BCQA*" format for multitask learning.

The above experimental results demonstrate that not only does AQF-RQ generate better pseudo-data, but the "*BQCA*" format is also more data-efficient and robust for lifelong language learning.

5.3 Generation for Different Domains

We conducted experiments on five sequence generation tasks for different domains by comparing LAMOL$_T$, LAMOL$_G$, and L2KD with AQF-RQ. The result is similar to Section 5.2.1 and is detailed in Appendix C. AQF-RQ is not only better than LAMOL-based baselines on performance but can also forward/backward transfer more knowledge among the same task but different domains.

5.4 Effectiveness of the Ask Question First

To verify the effectiveness of our proposed AQF, we experimentally apply AQF to each baseline: LAMOL$_T$, LAMOL$_G$, LAMOL$_R$, and L2KD. In Section 3.2, we stated that AQF is used better when the pseudo data is sufficient. Therefore, we set $\gamma = 0.2$ for experiments. This value is also the best setting for those baselines.

As shown in the Table 5, applying AQF to each baseline resulted in varying degrees of improvement. LAMOL$_T$ has the most noticeable improvement, with a 2 percentage point increase. This demonstrates that having the *A* close to the *C* using the "*BQCA*" format proposed by AQF can reduce information attenuation caused by the *A* being far away from the *C*. For example, the *C* and the *A* are separated by the *Q* in the "*BCQA*" format. This benefit is more obvious when the task boundary is clear (i.e. the sentence starts with [TASK]). [TASK] can tighten the bond between "*BQC*" and "*A*".

Applying AQF improves LAMOL$_G$ by 1.2 percentage points, which is less than LAMOL $_T$. It

is because the generated pseudo data is still biased toward new tasks due to the unclear task boundary (the sentence begins with the task-independent token [GEN]). However, thanks to the *A* being next to the *C* in the "*BQCA*" format, it is still capable of learning from a small amount of pseudo data.

For LAMOL$_R$, AQF can enhance the model by 1.4 percentage points, which is only 0.1% behind multi-task learning. LAMOL$_R$ uses the real data of the old task instead of the generated pseudo data when learning new tasks, so this further demonstrates that the "*BQCA*"format proposed by AQF is more conducive to lifelong language learning: it can use *Q* to help the model pay more attention to the important information in *C*, then further makes the generation of *A* more accurate.

For L2KD, the application of AQF improves the model by one percentage point, demonstrating that AQF is still applicable to the previous LAMOL-based enhanced approaches and is capable of effectively applied to other LAMOL-based studies.

For MTL, AQF can bring an improvement of 1 percentage point. At the same time, the improvement of each baseline demonstrates that the improvement brought by AQF is comprehensive: it is not only conducive to alleviating catastrophic forgetting but also improves the upper bound of the model. Therefore, we believe that AQF has the potential to apply LLL to real-world scenarios.

6 Effectiveness of the Replay Questions

Since the Replay QuestionsRQ is proposed based on AQF, we verify and analyze the effectiveness of the RQ based on Table 3 and 5. The AQF-RQ performance in Table 3 subtracts the corresponding LAMOL+AQF performance in Table 5 is the gains obtained by AQF-RQ come from

the RQ step. Therefore, $RQ_G^{0,2}$ = AQF-$RQ_G^{0,2}$ - LAMOL+$AQF_G^{0,2}$ = 1.1, and $RQ_T^{0,2}$ = AQF-$RQ_T^{0,2}$ - LAMOL+$AQF_T^{0,2}$ = 0.7. It demonstrates that the RQ step based on the AQF can help the model generate more task-corresponding pseudo samples by making $P(Q_i|B_i) > P(Q_{j\neq i}|B_i)$.

7 Conclusion and Future Work

This work proposed AQF-RQ which is a simple yet efficient and robust lifelong language learning method. We propose a new question-first data format that is train-efficient without additional computational resources and new parameters. In AQF's "*BQCA*" format, generating Q first makes it easier to generate task-corresponding pseudo-data, and A is more accurate because A is next to C. In addition, RQ can strengthen the model's attention to the problem, so that the model has a greater probability of generating task-corresponding questions, and further generates more task-corresponding pseudo-data. AQF-RQ effectively alleviates catastrophic forgetting, only 0.36% lower than multi-task learning. Due to a lack of computing resources, we did not experiment on larger datasets and longer task orders, which we leave as future work. In addition, we will investigate ways to enhance performance when the task boundary is unclear.

References

Leanne Bowler. 2010. Talk as a metacognitive strategy during the information search process of adolescents. *Inf. Res.*, 15(4).

David W. Braithwaite and Lauren Sprague. 2021. Conceptual knowledge, procedural knowledge, and metacognition in routine and nonroutine problem solving. *Cogn. Sci.*, 45(10).

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2020. Invariant rationalization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR.

Zhiyuan Chen and Bing Liu. 2018. *Lifelong Machine Learning, Second Edition*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

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, Online. Association for Computational Linguistics.

John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34:906–911.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

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.

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 *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Yufan Huang, Yanzhe Zhang, Jiaao Chen, Xuezhi Wang, and Diyi Yang. 2021. Continual learning for text classification with information disentanglement based regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2736–2746. Association for Computational Linguistics.

Kasidis Kanwatchara, Thanapapas Horsuwan, Piyawat Lertvittayakumjorn, Boonserm Kijsirikul, and Peerapon Vateekul. 2021. Rational LAMOL: A rationale-based lifelong learning framework. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2942–2953. Association for Computational Linguistics.

Tianlin Liu, Lyle Ungar, and João Sedoc. 2019. Continual learning for sentence representations using conceptors. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3274–3279. Association for Computational Linguistics.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7452–7467. Association for Computational Linguistics.

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*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Fei Mi, Liangwei Chen, Mengjie Zhao, Minlie Huang, and Boi Faltings. 2020. Continual learning for natural language generation in task-oriented dialog systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3461–3474. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 201–206. Association for Computational Linguistics.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mark B. Ring. 1997. CHILD: A first step towards continual learning. *Mach. Learn.*, 28(1):77–104.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning. *Proceedings of the ICLR 2020*.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2020. Distill and replay for continual language learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3569–3579, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Han Wang, Ruiliu Fu, Chengzhang Li, Xuejun Zhang, Jun Zhou, Xing Bai, Yonghong Yan, and Qingwei Zhao. 2022. Reminding the incremental language model via data-free self-distillation. *Applied Intelligence*, pages 1–23.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1711–1721. The Association for Computational Linguistics.

hyperparameter	value
optimizer	AdamW
adam epsilon	1×10^{-4}
learning rate	1×10^{-4}
weight of RQ task	$\eta = 0.2$
weight of LM task	$\lambda = 0.25$
max gradient norm	1.0
learning rate schedule	warmup linear
warmup ratio	0.005
max length	1024
top-k sampling	k=20

Table 6: The main hyperparameters in our experiment.

A   Datasets

As we can not obtain test set of SQuAD since it is hidden from the host, we use development set for testing. For other tasks, we use the corresponding test set. A normalized F1 (nF1) metric that lower text and remove punctuation and articles, is used to evaluate SQuAD and SRL. The exact match (EM) is used to evaluate SST. The turn-based dialogue state EM (dsEM) is used for WOZ. The ROUGE is used to evaluate E2ENLG and RNNLG (rest/hotel/tv/laptop)The size of each dataset is detailed in Table 7.

As shown in Table 7, SST, WOZ, E2ENLG and RNNLG (rest/hotel/tv/laptop) only have one question. The SQuAD has mainly 12 types of questions and other countless types of questions. SRL has 6 types of questions.

B   Hyperparameter

The main hyperparameters in our experiments are detailed in Table 6.

C   Sequence Generation for Different Domains

We conducted experiments on five sequence generation tasks for different domains by comparing LAMOL_T, LAMOL_G, and L2KD with AQF-RQ. Experiments are conducted with sufficient pseudo data (i.e. γ=0.2). Following L2KD, we select the task order from hard to easy: E2ENLG-RNNLG (rest) - RNNLG (hotel) - RNNLG (tv) - RNNLG (laptop). As shown in the Fig.3, we can observe that all models perform well except for Finetune and LAMOL_G. However, AQF-RQ can still outperform other baselines. When learning the same

Dataset	#Train	#Test	Metric	Question Type	Question
different types of tasks					
SQuAD	87599	10570	nF1	12+∞	what/who/whose/whom/when/ where/how/why/which/if/do/is … + other countless types of questions
SST	6920	1821	EM	1	is this review negative or positive?
SRL	6414	2201	nF1	6	what/who/whose/whom/ when/where/how/why …
WOZ	2536	1646	dsEM	1	what is the change in state?
suquence generation for different domains					
E2ENLG	6000	2000			
RNNLG (rest)	6228	1039			
RNNLG (hotel)	6446	1075	ROUGE	1	what is the natural language form?
RNNLG (tv)	8442	1407			
RNNLG (laptop)	7944	2649			

Table 7: The summarized results for datasets.

type of tasks in different domains, first we expect that previous tasks can promote the learning of new tasks. This is known as the forward transfer. In terms of the forward transfer, AQF-RQ is better at acquiring knowledge from previous tasks that is beneficial to new tasks. The second is backward transfer, which refers to learning new tasks while consolidating and strengthening previous tasks. It can be observed that AQF-RQ also has better backward transfer. For example, from the fourth task RNNLG.tv on, the backward transfer of AQF-RQ on the previous tasks is higher than other baselines. As the number of learned tasks increases, the unclear task boundary gradually becomes significantly weaker than the clear task boundary. For LAMOL$_G$, the performance on E2ENLG begins to drop significantly from the fourth task RNNLG.tv. But our AQF-RQ$_G$ remains on an upward trend. According to the above analysis, it can be concluded that AQF-RQ has stronger forward transfer and backward transfer capabilities in different domains of the same type of task.
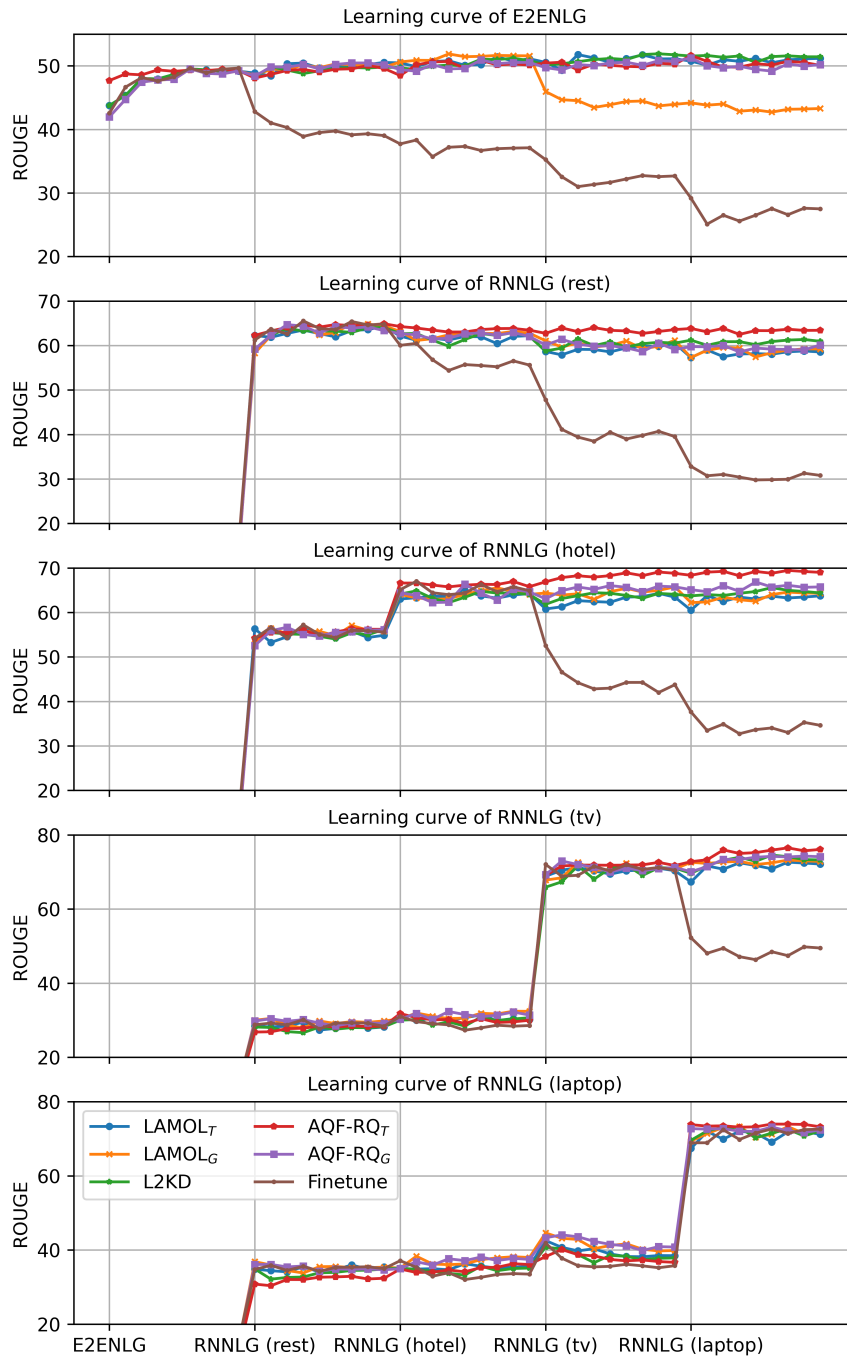
Figure 3: Results on sequence generation tasks for five different domains when $\gamma = 0.2$.