MKT: A Multi-Stage Knowledge Transfer Framework to Mitigate Catastrophic Forgetting in Multi-Domain Chinese Spelling Correction

Anonymous ACL submission

Abstract

Chinese Spelling Correction (CSC) aims to detect and correct spelling errors in given sentences. Recently, multi-domain CSC has gradually attracted the attention of researchers because it is more practicable. In this paper, we focus on the key flaw of the CSC model when adapting to multi-domain scenarios: the tendency to forget previously acquired knowledge upon learning new domain-specific knowledge (i.e., **catastrophic forgetting**). To address this, we propose a novel model-agnostic **M**ulti-stage **K**nowledge **T**ransfer (**MKT**) framework with an evolving teacher model and dynamic distillation weights for knowledge transfer in each domain, rather than focusing solely on new domain knowledge. It deserves to be mentioned that we are the first to apply continual learning methods to the multi-domain CSC task. Experiments [1] prove our method's effectiveness over traditional approaches, highlighting the importance of overcoming catastrophic forgetting to enhance model performance.

1 Introduction

Chinese Spelling Correction (CSC) plays a critical role in detecting and correcting spelling errors in Chinese text (Li et al., 2022c; Ma et al., 2022), enhancing the accuracy of technologies like Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) (Afli et al., 2016; Wang et al., 2018). In search engines, for example, CSC reduces human error, ensuring that users find the information they seek accurately.

In real applications, the input text may come from various domains, demanding that the model contains different domain-specific knowledge. As illustrated in Table 1, the word "强基(Strong Foundation)" is evidently common in the Chinese Education domain. Accurately correcting "张(open)" to "强(Strong)" requires the model to have specific knowledge about the Chinese Education domain. Therefore, some works have begun to focus on the impact of domain knowledge on the performance of CSC models (Lv et al., 2023a; Wu et al., 2023).

Input	他通过了张(zhāng)基计划。 He passed the Open Foundation plan.
+EDU	他通过了强(qiáng)基计划。 He passed the Strong Foundation plan.
+CHEM	他通过了羟(qiǎng) 基计划。 He passed the Hydroxyl project.
Target	他通过了强(qiáng)基计划。 He passed the Strong Foundation plan.

Table 1: Case of catastrophic forgetting in multi-domain CSC. red represents the misspelled character and blue represents the corrected character.

Previous works place greater emphasis on a model's ability to generalize to unseen domains, known as zero-shot performance, leveraging shared knowledge across different domains for generalization (Liu et al., 2023). However, in practical scenarios, for different domains, text correction needs often exist simultaneously and may evolve and increase over time. Therefore, CSC models must continuously learn and adapt across multiple domains. *This is not merely a problem of domain adaptation but a challenge of sequential learning and knowledge updating across multiple domains.* This aligns with the widely studied continual learning. **Hence, in this paper, we first incorporate the continual learning setting into CSC models.**

The core challenge of the continual learning setting is to minimize catastrophic forgetting of previously acquired knowledge while learning in new domains (Wang et al., 2024). As demonstrated in Table 1, when a CSC model learns educational-specific knowledge, it accurately corrects the word "强基(Strong Foundation)". However, after it continues to learn knowledge from the chemistry do-

[1] Our codes and data will be public after peer review.

main, it would learn the new knowledge of "羟基(hydroxyl)", but forget the education word "强基(Strong Foundation)". *Unfortunately, in previous multi-domain CSC studies, the challenge of this catastrophic forgetting of domain-specific knowledge has not been fully explored.*

In the field of Computer Vision, extensive studies are conducted on continual learning (Simon et al., 2022). We conduct experiments using common methods such as replay and knowledge distillation (Gou et al., 2021). These methods do help mitigate catastrophic forgetting to some extent, but there are still issues that need to be addressed, such as data imbalance and difficulties in updating the teacher model. To further improve upon this, we propose a novel model-agnostic Multi-stage Knowledge Transfer framework featuring an evolving teacher model. At each stage, the teacher model transfers its accumulated knowledge to the current student model, with distillation weights dynamically adjusting based on the data ratio. Finally, through extensive experiments and analyses, we demonstrate the effectiveness of our proposed method. The experimental results are shown in Table 3 and will be discussed in detail in Section 4.4. Our contributions are summarized as follows:

1. We are the first to highlight the catastrophic forgetting phenomenon of domain-specific knowledge in multi-domain CSC, a key challenge that must be overcome for CSC models to truly adapt to real multi-domain scenarios.

2. We present a model-agnostic MKT framework with an continuously evolving teacher model and dynamic distillation weights that effectively collaborate to mitigate domain-specific knowledge forgetting.

3. We conduct extensive experiments and thorough analyses to validate the effectiveness and competitiveness of our proposed method compared to other continual learning methods.

2 Related Work

2.1 Chinese Spelling Correction

In the field of CSC, we witness significant advancements in various model architectures and modules, as evidenced by recent works (Li et al., 2022b, 2023b; Zhang et al., 2023; Ye et al., 2023b, 2022; Ma et al., 2023; Ye et al., 2023a; Huang et al., 2023; Li et al., 2023d). Early models such as the Confusionset-guided Pointer Networks focus on optimizing at the dataset level by leveraging confusion sets for character generation. This technique enhances accuracy by considering commonly confused characters (Wang et al., 2019). Innovations in embeddings, like the REALISE model, improve model inputs by integrating semantic, phonetic, and visual information into character embeddings, thereby enriching the representational capacity of the model (Xu et al., 2021). Improvements in encoders are highlighted by models such as Soft-Masked BERT, which employs Soft MASK techniques post-detection to blend input characters with [MASK] embeddings. This method is effective for error prediction and has shown significant improvements in performance (Zhang et al., 2020). Another notable model, SpellGCN, constructs a character graph and maps it to interdependent detection classifiers based on BERT-extracted representations, showcasing innovative uses of graph neural networks in spelling correction (Cheng et al., 2020).

Previous research in multi-domain CSC emphasizes cross-domain knowledge sharing and generalization (Lv et al., 2023a). Typically, this involves training models on high-quality datasets to generalize effectively to specific domains. However, domain-specific knowledge is hard to generalize, and fine-tuning on multiple datasets can lead to catastrophic forgetting, where new knowledge overwrites old knowledge. This paper addresses catastrophic forgetting by introducing mechanisms that balance retaining existing knowledge with integrating new information. We propose a framework that mitigates forgetting while ensuring robust performance across multiple domains.

2.2 Continual Learning

In the field of continual learning, core strategies such as replay, regularization, and parameter isolation play pivotal roles (Liu et al., 2022; Li et al., 2022a; Wang et al., 2023; Dong et al., 2023; Li et al., 2023c). Replay methods, including techniques like GEM and MER, work by retaining training samples and using constraints or meta-learning to align gradients effectively (Lopez-Paz and Ranzato, 2017; Riemer et al., 2018). Regularization strategies, with Elastic Weight Consolidation (EWC) being a prime example, focus on preserving task-specific knowledge by emphasizing the importance of parameters that are critical to previous tasks (Kirkpatrick et al., 2017). Knowledge

distillation is another key approach, aiming at incremental training by transferring insights from larger models to smaller ones, thereby facilitating the integration of new knowledge while retaining old knowledge (Gou et al., 2021). Parameter isolation techniques, such as CL-plugin, address the issue of task interference by allocating unique parameters to different tasks, thus reducing the likelihood of overlap and interference (Ke et al., 2022).

Our work is pioneering in that it introduces continual learning to the multi-domain CSC task for the first time. Our MKT framework stands out as a model-agnostic approach, capable of being applied across various CSC models. By leveraging the strengths of existing continual learning strategies and integrating them into a cohesive framework, we aim to effectively mitigate catastrophic forgetting and enhance the adaptability of CSC models in multi-domain scenarios.

3 Our Approach

Our approach incorporates two dynamic mechanisms, designed for scenarios involving continual training across multiple domains. The primary mechanism features an evolving teacher model that continuously updates its knowledge base to encompass the most crucial knowledge from the previously trained domains. The secondary mechanism involves dynamic distillation weights, which better balance the loss between the teacher and student model. This provides a simple and effective solution for continual learning in multi-domain CSC.

3.1 Problem Formulation

The CSC task is to detect and correct spelling errors in Chinese texts. Given a misspelled sentence $X = \{x_1,\ x_2, ...,\ x_n\}$ with n characters, a CSC model takes X as input, detects possible spelling errors at character level, and outputs a corresponding correct sentence $Y = \{y_1,\ y_2, ...,\ y_n\}$ of equal length. This task can be viewed as a conditional sequence generation problem that models the probability of $p(Y|X)$. In multi-domain CSC tasks, assuming that there are n domains $D = \{D_1,\ D_2, ...,\ D_n\}$, these domains are trained sequentially, where each domain D_k is trained without access to the data from previous domains, from D_1 to D_{k-1}. Furthermore, after training domain D_k, we should consider the performance of all domains from D_1 to D_k, a metric which we will introduce in Section 4.1.

3.2 Structure of MKT framework

To tackle catastrophic forgetting, an intuitive solution is to transfer the knowledge previously acquired to the most recent model. The foundational idea revolves around transferring previously acquired knowledge to the latest model iteration. However, maintaining a distinct model for each stage quickly becomes untenable due to escalating storage and computational requirements with the addition of each domain.

Using the concept of knowledge distillation, if a fixed teacher model is used to distill knowledge into each domain-specific student model, it remains a challenging problem to ensure that the student model learns the important knowledge from all previously learned domains.

To address this challenge, our framework employs a dynamic teacher model strategy. As illustrated in Figure 1, this teacher model acts as a comprehensive knowledge repository, effectively serving as a backup of the student model from the previous stage to calculate the distillation loss for the current stage's student model. It encapsulates all the domain-specific knowledge accumulated to date, providing crucial guidance for the model training in the current phase. Additionally, we conduct experiments to explore how to a priori select appropriate distillation weights (experimental results are shown in Table 4), so that framework can dynamically adjusted distillation weights during training to achieve better performance.

3.3 MKT framework for Multi-domain CSC

We consider the scenario where the training is comprised of m stages, denoted by $k = 1, 2..., m$. At k-th stage, a subset of data $\{x_k^{(i)},\ y_k^{(i)}\}_{i=1}^{T_k}$ are fed to the model, where T_k refers to the number of samples at k-th stage, $x_k^{(i)}$ refers to i-th sample at k-th stage.

Assume that $u_k(\cdot)$ is an unknown target function that maps each $x_k^{(i)}$ to $y_k^{(i)}$ at stage k, i.e., $y_k^{(i)} = u_k(x_k^{(i)})$. Under the continual learning setting, our goal is to train a CSC model $g(\cdot\,; w)$ parameterized by w, such that $g(\cdot\,; w)$ not only fits well to $u_k(\cdot)$, but also fits $u_{k-1}(\cdot)\ , u_{k-2}(\cdot), \cdots,\ u_1(\cdot)$ in early stages to alleviate catastrophic forgetting.

We need to minimize the loss function to optimize the model weights:

$$L^{(k)} = \lambda L_s^{(k)} + L_h^{(k)}. \tag{1}$$

In the equation, λ is a hyper-parameter that ranges

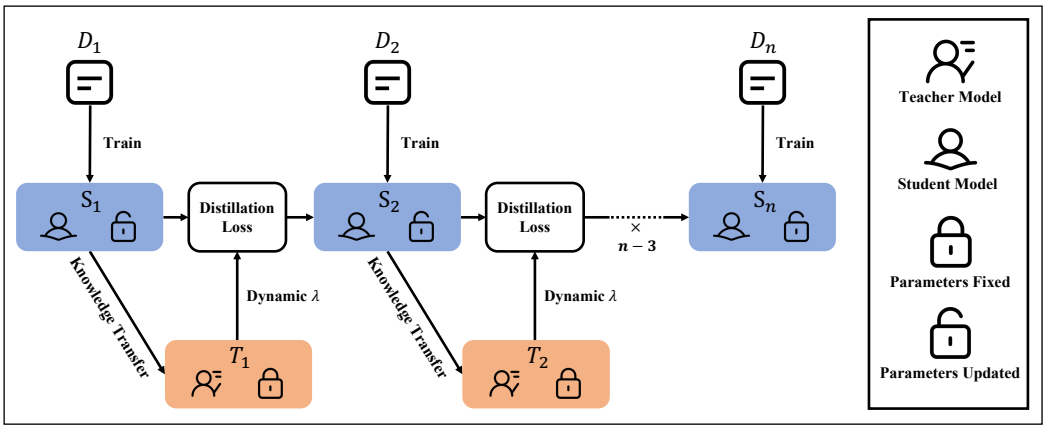

Figure 1: Overview of the MKT framework and the pipeline for multi-domain training.

from $[0, 1]$. $L_s^{(k)}$ is the knowledge distillation loss, calculating cross entropy between the output probabilities of teacher model $g(\cdot\,; w_{k-1})$ and student model $g(\cdot\,; w_k)$:

$$L_s^{(k)} = -\sum_{i=1}^{T_k} g(x_k^{(i)};\, \omega_{k-1}) \times \log g(x_k^{(i)};\, \omega_k). \tag{2}$$

$L_h^{(k)}$ is the cross-entropy loss between the output of student model $g(\cdot\,; w_k)$ and ground truth y_k:

$$L_h^{(k)} = -\sum_{i=1}^{T_k} y_k^{(i)} \times \log g(x_k^{(i)};\, \omega_k). \tag{3}$$

The choice of λ is related to the ratio of domain data and old data, S_d is domain data scale and S_o is old data scale:

$$\lambda = \frac{S_d}{S_o} \tag{4}$$

Algorithm 1 MKT Framework

Input: Training set D_k, Student model S_{k-1}
Output: Student model S_k

1: Copy S_{k-1} as the teacher model T_k
2: Freeze the parameters of T_k
3: Calculate λ according to Equation4.
4: S_k forward propagation and calculates the loss guided by T_k according to Equation 1
5: Optimize the parameters of S_k
6: **Return** S_k

As shown in Algorithm 1, during the training of the k-th domain, we employ the model refined from the preceding $k-1$ domains (i.e, S_{k-1}), as the teacher model T_k, alongside the concurrently trained student model S_k. The parameters of T_k are frozen. The final loss is the dynamically weighted summation of the knowledge distillation loss $L_s^{(k)}$ and the original CSC task loss $L_h^{(k)}$, with the weights as shown in Equation 4.

4 Experiment and Result

In this section, we introduce our multi-domain datasets and experiments, aiming to validate the superiority of our proposed MKT framework compared to other methods in mitigating catastrophic forgetting and enhancing generalization capability.

4.1 Datasets and Metrics

Training Set	Domain	Sent	Avg.Length	Errors
Wang271K	General	271,329	42.6	381,962
SIGHAN13	General	700	41.8	343
SIGHAN14	General	3,437	49.6	5,122
SIGHAN15	General	2,338	31.1	3,037
CAR	CAR	2,743	43.4	1,628
MED	MED	3,000	50.2	2,260
LAW	LAW	1,960	30.7	1,681
Test Set	**Domain**	**Sent**	**Avg.Length**	**Errors**
SIGHAN15	General	1,100	30.6	703
CAR	CAR	500	43.7	281
MED	MED	500	49.6	356
LAW	LAW	500	29.7	390

Table 2: Statistics of the datasets we use.

Considering the multi-domain setting we focus on, we set up four domains, namely **General, Car, Medical, and Legal** domains. The reason for this setting is that the differences in characteristics between these domains are the most obvious, which brings the most serious catastrophic forgetting to CSC models. For the general domain, as in previous work, we also use SIGHAN13/14/15 (Wu et al.,

2013; Yu and Li, 2014; Tseng et al., 2015) and Wang271K (Wang et al., 2018) as training data and SIGHAN15 test set as our test data. For other special domains, we utilize the data resources released by LEMON (Wu et al., 2023) and ECSpell (Lv et al., 2023b), and randomly take 500 samples from the original data of each domain as the test set. The dataset statistics are presented in the Table 2.

Our evaluation predominantly relies on the sentence-level F1 score, a widely acknowledged metric. This criterion is notably stringent, adjudging a sentence as accurate solely when every error within is precisely identified and rectified, thereby providing a more rigorous evaluation compared to character-level metrics. In each table, Avg represents the overall performance after training on all domains. Unlike average accuracy (AA) (Wang et al., 2023), we use the average sentence-level F1 score, which is a more stringent metric than AA.

4.2 Baseline Methods

To validate the model-agnostic nature of MKT, we selected three widely used CSC baselines with different architectures to evaluate the effectiveness of our approach across various structures:

1. ***BERT*** (Devlin et al., 2019): Directly fine-tune the *chinese-roberta-wwm-ext* model using a series of domain-specific datasets.

2. ***Soft-Masked BERT*** (Zhang et al., 2020): Incorporates a soft masking process after the detection phase, where it calculates the weighted sum of the input and [MASK] embeddings.

3. ***REALISE*** (Xu et al., 2021): Models semantic, phonetic and visual information of input characters, and selectively mixes information in these modalities to predict final corrections.

To validate the effectiveness of our MKT framework, we compare it with different continual learning methods on the aforementioned models to demonstrate the superiority of our approach:

1. ***Joint-Training*** (Caruana, 1997): Mix the new domain data with the old data for training.

2. ***Fine-tuning***: Without using anti-forgetting methods, the model simply fine-tune on a series of domain data.

3. ***Replay(random)*** (Chaudhry et al., 2019): Randomly sample 1% from the old data and combine it with the new data for training.

4. ***Replay(RAP)*** (Li et al., 2024): **R**eplay **A**ccording im**P**ortance RAP selects 10% of the old data based on importance (i.e., the loss of the old data on the old model) and mixes it with the new domain data for training.

4.3 Implementation Details

In the main experiment, we initially train the models on General dataset, which consists of Wang271K combined with double the amount of SIGHAN data. This is followed by training on the CAR, MED, and LAW datasets using various continual learning methods, including Joint-Training, fine-tuning, replay (random), and replay (RAP). Upon completion of training, we evaluate the performance of the final model across all domain-specific datasets to gauge its effectiveness.

Additionally, auxiliary experiments are conducted using our top-performing REALISE model. These experiments investigate several factors such as determining the optimal λ, assessing the appropriate buffer size, examining the effects of different training orders, and performing ablation studies to understand the contribution of each component.

For all experiments, we train the aforementioned datasets for 10 epochs with a batch size of 64. The learning rates are 5e-5 for REALISE and BERT models, and 1e-4 for the Soft-Masked BERT model. Our approach incorporates a knowledge transfer process at each domain, where the λ between L_h and L_s is updated prior to training each domain according to Equation 4. The hyper parameter settings for the auxiliary experiments remain consistent with those used in our main experiments.

4.4 Results and Analyses

Main Results From Table 3, we see that after the optimization of our MKT, whether it is BERT, Soft-Masked BERT specially designed for CSC, or REALISE that integrates multi-modal information, their performance improves in all domains. This reflects the effectiveness and the model-agnostic characteristic of our proposed MKT framework.

Regarding the comparison between MKT and other continual learning methods, it can be observed that Joint-Training, due to the General dataset being much larger than the specific domain datasets, can effectively mitigate forgetting in the General dataset but fails to adequately learn the new dataset's knowledge. Fine-tuning, without any measures against forgetting, results in significant loss of previously acquired knowledge. Randomly

Model	Method	General	CAR	MED	LAW	Avg
BERT	Joint-Training	71.50	31.75	42.58	60.41	51.56
	Fine-tuning	67.41	33.50	42.86	62.35	51.53
	+Replay(random)	70.07	34.87	41.33	59.51	51.45
	+Replay(RAP)	70.09	36.22	43.00	58.25	51.89
	+MKT(Ours)	68.58	36.18	43.56	62.47	**52.70**$^{\uparrow}$
Soft-Masked BERT	Joint-Training	60.96	26.67	40.00	58.19	46.46
	Fine-tuning	54.22	30.73	43.88	68.54	49.34
	+Replay(random)	47.45	23.88	39.51	64.30	43.79
	+Replay(RAP)	54.48	22.86	46.52	60.59	46.11
	+MKT(Ours)	60.90	35.64	52.21	70.40	**54.79**$^{\uparrow}$
REALISE	Joint-Training	76.77	26.81	50.72	68.72	55.76
	Fine-tuning	70.78	27.48	53.33	70.59	55.55
	+Replay(random)	75.78	27.83	53.81	69.25	56.67
	+Replay(RAP)	76.10	31.51	50.33	69.76	56.93
	+MKT(Ours)	73.84	31.25	54.1	70.18	**57.34**$^{\uparrow}$

Table 3: Performance on the test set of each domain after training on all datasets.

selecting old data to train together with the new data shows relatively good performance when the data is more balanced. Selecting old data based on importance and training it together with new domain data achieve better results compared to other methods. It performs only slightly worse than our MKT framework. However, its training time is ten times longer than our method.

Parameter Study To explore the impact of the key parameter λ, we conduct experiments with different λ values on REALISE + MKT under varying ratios of new domain and old data. As shown in Table 4, we perform experiments on the General dataset and the subsequent three specific domain datasets, selecting a portion of the data from the General dataset as the old dataset. The size of the old dataset was set to be 50, 20, and 10 times that of the corresponding specific domain data. When λ was set to 0.5, 1, and 2 times the ratio of the domain dataset size to the old dataset size, the results generally showed stable improvements over the baseline (i.e., $\lambda = 0$). In particular, when λ matched the ratio of the domain data to the old data, it perform best across all domains.

Therefore, intuitively, for MKT, it can choose the appropriate λ based on the ratio of the new domain data to the old data to achieve optimal performance.

Buffer study Due to the severe imbalance in the scale of new domain and old data in the Joint-Training method, we explore the optimal buffer size for the replay method by conducting a series of experiments on REALISE. As shown in Table 5, when randomly selecting the buffer, the best performance is achieved by choosing 1% of the old data as the buffer size, because the size of the new and old data is relatively balanced. However, when selecting the buffer based on sample importance, choosing 10% of the old data as the buffer size yields the best results, even though it consumes a significant amount of training time. This is because selecting important samples for training allows for better learning of the most critical knowledge from both the new and old data.

Catastrophic Forgetting The above analysis convincingly demonstrate that the MKT framework outperforms other continual learning methods in overall performance after training across all domains. To better observe the forgetting at each stage when training on subsequent domain datasets, we selecte the best-performing model from Table 3 (i.e., REALISE) and examine its performance loss (i.e., catastrophic forgetting) on the General dataset after incremental training with data from other domains, as shown in Figure 2.

The performance loss of REALISE on the Gen-

$\frac{S_d}{S_o}$	λ	CAR			MED			LAW		
		General	**Domain**	**Avg**	**General**	**Domain**	**Avg**	**General**	**Domain**	**Avg**
0.02	0	66.73	30.42	48.58	65.65	47.17	56.41	66.07	59.45	62.76
	0.01	66.97	30.25	48.61	65.78	47.60	56.69	66.9	58.27	62.59
	0.02	67.26	30.96	**49.11**$^{\uparrow}$	66.67	47.27	**56.97**$^{\uparrow}$	66.91	59.01	**62.96**$^{\uparrow}$
	0.04	67.51	30.19	48.85	67.17	45.16	56.17	66.84	58.38	62.61
0.05	0	62.25	29.85	46.05	62.36	42.44	52.40	61.27	58.5	59.89
	0.025	64.10	29.42	46.76	61.84	43.89	52.87	61.87	59.84	60.86
	0.05	65.80	30.17	**47.99**$^{\uparrow}$	63.82	41.96	**52.89**$^{\uparrow}$	62.30	60.63	**61.47**$^{\uparrow}$
	0.1	64.85	30.17	47.51	62.94	41.27	52.11	63.00	57.82	60.41
0.1	0	55.07	27.44	41.26	57.69	40.18	48.94	53.68	54.94	54.31
	0.05	58.46	28.03	43.25	57.48	41.91	49.70	54.53	55.62	55.08
	0.1	59.13	28.51	**43.82**$^{\uparrow}$	58.81	41.38	**50.10**$^{\uparrow}$	55.66	55.20	**55.43**$^{\uparrow}$
	0.2	57.47	25.70	41.59	58.96	35.70	47.33	55.60	54.22	54.91

Table 4: Selection of optimal distillation weights (λ) under different domain(S_d) and old(S_o) data ratios.

Model	Method	Buffer size	General	CAR	MED	LAW	Avg
REALISE	Replay(random)	0.001	74.14	27.56	54.07	67.88	55.91
		0.01	75.78	27.83	53.81	69.25	**56.67**$^{\uparrow}$
		0.1	74.44	30.33	51.94	67.87	56.15
	Replay(RAP)	0.001	74.31	26.77	49.37	67.88	54.58
		0.01	75.48	31.51	48.75	68.67	56.10
		0.1	76.10	31.51	50.33	69.76	**56.93**$^{\uparrow}$

Table 5: Performance of different replay methods and various buffer sizes.

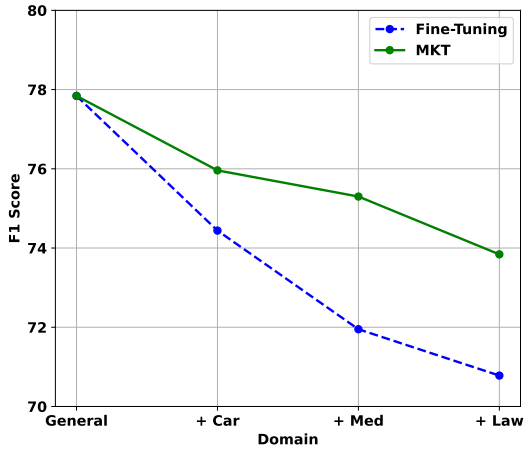

Figure 2: The phenomenon of model forgetting General-domain knowledge during incremental domain training.

eral dataset is much smoother when optimized with MKT, indicating that MKT framework effectively mitigates catastrophic forgetting at each stage.

Training Order To investigate whether our MKT framework can mitigate catastrophic forgetting across different training orders, we also experiment with alternative sequences. For instance, we choose the best-performing model, REALISE, from Table 3 and conduct training on it according to the new sequence, which is the reverse of the main experiment's order. As shown in Table 6, the table headers represent the training order of the domains, despite varying degrees of forgetting, MKT framework effectively mitigates catastrophic forgetting.

4.5 Ablation Study

MKT differs from knowledge distillation in two key aspects: a continuously evolving teacher model and distillation weights that dynamically change with the ratio of new and old data. To validate the effectiveness of these two mechanisms, we conduct ablation experiments on REALISE without these optimizations. As shown in Table 7, the two dynamic mechanisms of MKT effectively mitigate catastrophic forgetting, with the performance improvement brought by the dynamically evolving teacher being more significant.

In knowledge distillation, the teacher model is fixed. Although it contains extensive knowledge from the General dataset, it cannot be continually updated with subsequent domain knowledge. Con-

Model	Method	General	CAR	MED	LAW	Avg
REALISE	Knowledge distillation	74.23	29.69	52.68	67.61	56.05
	MKT(Ours)	73.84	31.25	54.10	70.18	**57.34**$^{\uparrow}$
Model	**Method**	**General**	**LAW**	**MED**	**CAR**	**Avg**
REALISE	Knowledge distillation	74.54	61.64	44.91	30.37	52.87
	MKT(Ours)	74.29	64.07	48.99	28.81	**54.04**$^{\uparrow}$

Table 6: The impact of training order on MKT Performance.

Model	Method	General	CAR	MED	LAW	Avg
REALISE	Knowledge distillation	74.23	29.69	52.68	67.61	56.05
	+ evolving teacher	72.74	29.25	55.28	70.85	57.03
	+ dynamic λ	74.13	30.30	53.74	67.34	56.38
	MKT(Ours)	73.84	31.25	54.10	70.18	**57.34**$^{\uparrow}$

Table 7: The impact of dynamic distillation weights (λ) and the evolving teacher model on performance.

sequently, while it effectively reduces forgetting in the General dataset, significant forgetting of previously learned domain knowledge still occurs after training on all domains.

A continuously evolving teacher model can incorporate the most important knowledge previously learned, effectively reducing the student's forgetting of prior knowledge. For dynamic distillation weights, we provided experimental results in Table 4. MKT's adaptation to the ratio of domain and old data allows it to better learn the most important knowledge from domain and old data. Using dynamic distillation weights alone can only provide limited performance improvement.

Our MKT combines these two dynamic mechanisms. While forgetting on the General dataset is slightly greater than with the fixed teacher method, overall anti-forgetting performance in subsequent domain learning significantly improves.

4.6 Case Study

To further verify the effectiveness of our MKT in mitigating catastrophic forgetting in multi-domain CSC, we present some cases in Table 8. For a test sentence in the CAR domain, REALISE accurately corrects errors after fine-tuning on CAR. However, after further fine-tuning on the MED domain, it can no longer correct successfully and instead predicts "氰(cyanide)" related to the medical domain. This is a typical catastrophic forgetting case where old domain knowledge is washed away by new domain knowledge. It can be seen that with the optimization of MKT, REALISE effectively avoids the occurrence of catastrophic forgetting.

Circumventing Catastrophic Forgetting	
Input	年轻人的青量级玩乐SUV
+CAR(Fine-tuning)	年轻人的轻量级玩乐SUV
+CAR(+MKT)	年轻人的轻量级玩乐SUV
+MED(Fine-tuning)	年轻人的氰量级玩乐SUV
+MED(+MKT)	年轻人的轻量级玩乐SUV
Target	年轻人的轻量级玩乐SUV

Table 8: Cases from the CAR test set, conducted on the REALISE model, show that the MKT framework mitigates over-correction and catastrophic forgetting.

5 Conclusion

This paper demonstrates through experimentation that existing CSC models, when adapting to multi-domain scenarios, tend to forget previously acquired domain-specific knowledge, a phenomenon known as catastrophic forgetting. Consequently, we propose an effective, model-agnostic MKT framework with an evolving teacher model and dynamic distillation weights to balance retaining existing knowledge with integrating new information, effectively mitigating catastrophic forgetting. Extensive experiments and detailed analyses highlight the significance of addressing catastrophic forgetting, proving the superiority of our method over other continual learning approaches.

Limitations

Our method specifically focuses on the Chinese language. However, other languages, such as English, could also benefit from our approach, and we plan to conduct related research on English contexts in the future. Additionally, we did not compare our proposed method with large language models (LLMs) commonly used in experiments. The primary reason is that representative LLMs still lag behind traditional fine-tuned smaller models in the CSC task, as has been demonstrated by some related works (Li et al., 2023a).

Of course, our main contribution is proposing a model-agnostic framework to mitigate catastrophic forgetting. We believe that combining the MKT framework with current LLMs could be a highly practical direction for future work.

References

Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 962–966, Portorož, Slovenia. European Language Resources Association (ELRA).

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, PuneetK. Dokania, PhilipH.S. Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning. *Cornell University - arXiv,Cornell University - arXiv*.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. A survey of natural language generation. *ACM Comput. Surv.*, 55(8):173:1–173:38.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11514–11525. Association for Computational Linguistics.

Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual training of language models for few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10216, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Jiyong Li, Dilshod Azizov, LI Yang, and Shangsong Liang. 2024. Contrastive continual learning with importance sampling and prototype-instance relation distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13554–13562.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Y. Li, F. Zhou, Haitao Zheng, and Qingyu Zhou. 2023a. On the (in)effectiveness of large language models for chinese text correction. *ArXiv*, abs/2307.09007.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023b. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Yinghui Li, Shulin Huang, Xinwei Zhang, Qingyu Zhou, Yangning Li, Ruiyang Liu, Yunbo Cao, Hai-Tao Zheng, and Ying Shen. 2023c. Automatic context pattern generation for entity set expansion. *IEEE Trans. Knowl. Data Eng.*, 35(12):12458–12469.

Yinghui Li, Yangning Li, Yuxin He, Tianyu Yu, Ying Shen, and Hai-Tao Zheng. 2022a. Contrastive learning with hard negative entities for entity set expansion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1077–1086. ACM.

Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Yangning Li, Shulin Huang, Ruiyang Liu, Chao Li, Yunbo Cao, and Haitao Zheng. 2022b. Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 238–249. Association for Computational Linguistics.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2023d. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. *CoRR*, abs/2311.11268.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022c. The past mistake is the future wisdom: Error-driven contrastive probability optimization for Chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3202–3213, Dublin, Ireland. Association for Computational Linguistics.

Linfeng Liu, Hongqiu Wu, and Hai Zhao. 2023. Chinese spelling correction as rephrasing language model.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7):100520.

David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qi Lv, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. 2023a. General and domain-adaptive chinese spelling check with error-consistent pretraining. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5).

Qi Lv, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. 2023b. General and domain-adaptive chinese spelling check with error-consistent pretraining. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–18.

Shirong Ma, Yinghui Li, Haojing Huang, Shulin Huang, Yangning Li, Hai-Tao Zheng, and Ying Shen. 2023. Progressive multi-task learning framework for chinese text error correction. *CoRR*, abs/2306.17447.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 576–589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. *ArXiv*, abs/1810.11910.

Christian Simon, Masoud Faraki, Yi-Hsuan Tsai, Xiang Yu, Samuel Schulter, Yumin Suh, Mehrtash Harandi, and Manmohan Chandraker. 2022. On generalizing beyond domains in cross-domain continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9265–9274.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37, Beijing, China. Association for Computational Linguistics.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2527, Brussels, Belgium. Association for Computational Linguistics.

Dingmin Wang, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for Chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5780–5785, Florence, Italy. Association for Computational Linguistics.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2023. A comprehensive survey of continual learning: Theory, method and application. *ArXiv*, abs/2302.00487.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for chinese spelling correction. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1:10743–10756.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at SIGHAN bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42, Nagoya, Japan. Asian Federation of Natural Language Processing.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 716–728, Online. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. 2023a. Mixedit: Revisiting data augmentation and beyond for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10161–10175. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Shirong Ma, Rui Xie, Wei Wu, and Hai-Tao Zheng. 2022. Focus is what you need for chinese grammatical error correction. *CoRR*, abs/2210.12692.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023b. CLEME: debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6174–6189. Association for Computational Linguistics.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 220–223.

Ding Zhang, Yinghui Li, Qingyu Zhou, Shirong Ma, Yangning Li, Yunbo Cao, and Hai-Tao Zheng. 2023. Contextual similarity is more valuable than character similarity: An empirical study for chinese spell checking. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890, Online. Association for Computational Linguistics.