Mitigating Catastrophic Forgetting in Multi-domain Chinese Spelling Correction by Multi-stage Knowledge Transfer Framework

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, which utilizes a continuously evolving teacher model 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 the effectiveness of our proposed method, and further analyses demonstrate the importance of overcoming catastrophic forgetting for improving the model performance.

1 Introduction

Chinese Spelling Correction (CSC) plays a critical role in detecting and correcting spelling errors in Chinese text (Li et al., 2022; 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 practical applications, the input text may 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 Chinese Education domain. Accurately correcting "张(open)" to "强(Strong)" requires the model to have specific knowledge about the Chinese Education domain. Therefore, some related works begin to focus on the impact of multi-domain knowledge on the performance of CSC models (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 CSC.

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, this paradigm falls short of enabling models to retain domain-specific knowledge, which is necessary for nuanced understanding and application. Human learning processes can continuously acquire new domain-specific knowledge without losing their learned old knowledge. Therefore, in this paper, we first investigate continual learning, which aligns perfectly with the human learning process, into CSC models for addressing this issue.

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 the educational-specific word "强基(Strong Foundation)", it accurately corrects errors. However, after it continues to learn knowledge from the chemistry domain, it would learn the new knowledge of "羟基(hydroxyl)", but forget the education word "强基(Strong Foundation)". However, in the previous works of multi-domain CSC, this catastrophic

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

forgetting challenge remains unexplored.

To mitigate catastrophic forgetting in multi-domain CSC, we devise a multi-stage knowledge transfer framework based on continual learning, which employs a dynamically evolving teacher model that at each stage imparts all its previously accumulated knowledge to the current student model. Finally, through extensive experiments and analysis, we demonstrate the effectiveness of our proposed method. Our contributions are summarized as follows:

1. We are the first to pay attention to the catastrophic forgetting phenomenon of multi-domain CSC, which is the key challenge that must be overcome for the CSC model to truly adapt to real multi-domain scenarios.

2. We present a model-agnostic MKT framework that leverages the idea of continual learning to significantly suppress catastrophic forgetting.

3. We conduct extensive experiments and solid analyses to verify the effectiveness and competitiveness of our proposed methods.

2 Our Approach

Our approach is a special form of knowledge distillation that takes into account scenarios involving multiple stages of training. We leverage the knowledge acquired from these stages. This strategy of multi-stage knowledge transfer provides an effective solution to the challenges encountered in continual learning.

2.1 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.

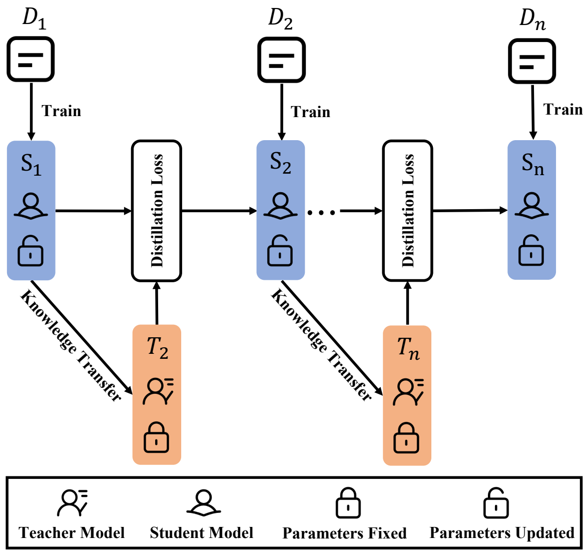

Figure 1: Overview of MKT framework.

2.2 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 stage.

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.

2.3 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:

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

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

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

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: S_k forward propagation and calculates the loss guided by T_k according to Equation 1
4: Optimize the parameters of S_k
5: **Return** S_k

As shown in Algorithm 1, during the training phase 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 a weighted summation of the knowledge distillation loss $L_s^{(k)}$ and the original loss of the CSC task $L_h^{(k)}$.

3 Experiment and Result

3.1 Datasets and Metrics

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., 2023), and randomly take 500 samples from the original data of each domain as the test set. The dataset statistics are presented in the Appendix B.

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.

3.2 Baseline Methods

We select three widely used CSC baselines that embody varying integration of sensory inputs, to assess the efficacy of our method in diverse structural contexts: ***BERT*** (Devlin et al., 2019) is to directly fine-tune the *chinese-roberta-wwm-ext* model with the training data. ***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. ***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. Other implementation details are shown in the Appendix C.

3.3 Results and Analyses

Main Results From Table 2, 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, performance improvements have been achieved in all domains. This reflects the effectiveness and the model-agnostic characteristic of our proposed MKT framework.

Parameter Study To explore the impact of the key parameter λ, we conduct experiments on BERT+MKT using varying λ values. As Table 3 indicates, settings λ between 0.005 and 0.02 stably bring improvements over the baseline. Particularly, setting λ at 0.01 performs best in all domains. We think that the main reason for this phenomenon is that the amount of training data in each special domain accounts for approximately 1% of the amount of general training data (as shown in Appendix B).

Backbone	Model	General	CAR	MED	LAW	Avg
BERT	Baseline	67.41	33.50	42.86	62.35	51.53
	+MKT(Ours)	67.90$^{\uparrow}$	35.86$^{\uparrow}$	43.46$^{\uparrow}$	62.88$^{\uparrow}$	52.53$^{\uparrow}$
Soft-Masked BERT	Baseline	54.22	30.73	43.88	68.54	49.34
	+MKT(Ours)	60.98$^{\uparrow}$	35.11$^{\uparrow}$	51.27$^{\uparrow}$	70.68$^{\uparrow}$	54.51$^{\uparrow}$
REALISE	Baseline	70.78	27.48	53.33	70.59	55.55
	+MKT(Ours)	72.74$^{\uparrow}$	29.25$^{\uparrow}$	55.28$^{\uparrow}$	70.85$^{\uparrow}$	57.03$^{\uparrow}$

Table 2: Performance on the test set of each domain after training on all datasets.

λ	General	CAR	MED	LAW	Avg
0	67.41	33.50	42.86	62.35	51.53
0.001	65.42	34.00$^{\uparrow}$	43.13$^{\uparrow}$	62.07	51.16
0.005	65.92	34.80$^{\uparrow}$	43.19$^{\uparrow}$	62.22	51.53$^{\uparrow}$
0.01	67.90$^{\uparrow}$	35.86$^{\uparrow}$	43.46$^{\uparrow}$	62.88$^{\uparrow}$	52.53$^{\uparrow}$
0.015	66.73	36.10$^{\uparrow}$	44.29$^{\uparrow}$	61.97	52.27$^{\uparrow}$
0.02	66.48	37.47$^{\uparrow}$	42.92$^{\uparrow}$	62.63$^{\uparrow}$	52.38$^{\uparrow}$
0.05	67.41	32.32	41.98	61.66	50.84
0.1	68.18$^{\uparrow}$	31.02	41.09	60.62	50.23
0.2	67.74$^{\uparrow}$	27.55	38.31	56.96	47.64
0.5	66.47	23.53	27.09	44.44	40.38
0.8	60.64	12.35	15.15	26.22	28.59

Table 3: Performance of BERT+MKT on each domain after training across all domains with different λ.

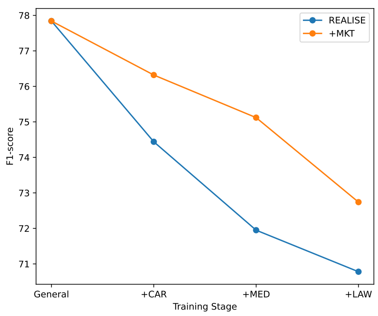

Figure 2: The phenomenon of model forgetting general-domain knowledge during incremental domain training.

Therefore, intuitively for MKT, an appropriate λ can be selected based on the ratio of general training data to training data in other domains to obtain optimal performance.

Catastrophic Forgetting As shown in Figure 2, we select the best-performing model (i.e., REALISE) in Table 2 to observe its performance loss (i.e. catastrophic forgetting) in the general domain after being incrementally trained with other domain data. Obviously, we see that after the optimization of MKT, the performance loss of REALISE is much smoother, which shows that catastrophic forgetting is well alleviated by our proposed MKT.

3.4 Case Study

Circumventing Catastrophic Forgetting	
Input	年轻人的青量级玩乐SUV
+CAR(REALISE)	年轻人的轻量级玩乐SUV
+CAR(+MKT)	年轻人的轻量级玩乐SUV
+MED(REALISE)	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 4: Cases from the CAR test set to show MKT mitigates over-correction and catastrophic forgetting.

To further verify the effectiveness of our MKT in mitigating catastrophic forgetting in multi-domain CSC, we present some cases in Table 4. As shown in table 4, for the test sentence in the CAR domain, when REALISE has just been trained on the CAR domain, it can accurately correct errors. However, when REALISE is then trained 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.

4 Conclusion

This paper demonstrates through experimentation that existing CSC models, when adapting to multi-domain scenarios, tend to forget previously acquired knowledge while learning new domain-specific information, a phenomenon known as catastrophic forgetting. Consequently, we propose an effective, model-agnostic framework for multi-stage knowledge transfer to mitigate catastrophic forgetting. Extensive experiments and detailed analyses demonstrate the importance of catastrophic forgetting we focus on and the effectiveness of our proposed method.

Limitations

We do not compare our proposed method against commonly used Large Language Models (LLMs) in our experiments. The primary reason is that in the CSC task, representative LLMs still lag behind traditional fine-tuned smaller models, which has been proved by many related works. In addition, our approach specifically focuses on the Chinese scenarios. However, other languages, such as English, could also benefit from our methodology. We will conduct related studies on English scenarios in the future.

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).

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.

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

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.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022. 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.

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. 2023. 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, 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.

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.

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.

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.

A Related Work

This section comprehensively reviews CSC research, structured according to the data flow within correction models, and also delves into three principal methods in continual learning.

A.1 Chinese Spelling Correction

In CSC, we witness significant advancements in various model architectures and modules.Early models like Confusionset-guided Pointer Networks optimize at the dataset level, leveraging confusion sets for character generation to enhance accuracy through commonly confused characters (Wang et al., 2019). Innovations in embeddings, such as REALISE, improve model inputs by integrating semantic, phonetic, and visual information into character embeddings (Xu et al., 2021). Encoder improvements are highlighted by Soft-Masked BERT, which employs Soft MASK techniques post-detection to blend input with [MASK] embeddings for effective error prediction (Zhang et al., 2020). SpellGCN innovatively constructs a character graph, mapping it to interdependent detection classifiers based on BERT-extracted representations (Cheng et al., 2020). While previous multi-domain CSC research emphasizes shared knowledge and generalization across domains, this paper pioneers in addressing the catastrophic forgetting of domain-specific knowledge.

A.2 Continual Learning

In continual learning, replay, regularization, and parameter isolation stand as core strategies (Wang et al., 2023). Replay methods like GEM and MER retain training samples, using constraints or meta-learning to align gradients (Lopez-Paz and Ranzato, 2017; Riemer et al., 2018). Regularization, exemplified by Elastic Weight Consolidation (EWC), focuses on preserving task-specific knowledge by prioritizing parameter importance (Kirkpatrick et al., 2017). Knowledge distillation aims at incremental training, transferring insights from larger to smaller models (Gou et al., 2021). Parameter isolation techniques, such as CL-plugin, allocate unique parameters to different tasks, reducing interference (Ke et al., 2022). Our work introduces continual learning to multi-domain CSC for the first time, with our MKT framework being model-agnostic across various CSC models.

B Statistics of the datasets

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	2,744	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 5: Statistics of the datasets, including the number of sentences, the average length of sentences in tokens, and the number of errors in characters.

C Implementation Details

For the three models, we initially train them on the general dataset, followed by successive training on the CAR, MED and LAW datasets. After the completion of training, we test the final models' performance on all datasets.

In the experiments, we train on the aforementioned datasets for 10 epochs each, with a batch size of 64 and a learning rate of 5e-5. The baseline method simply involved sequential training on the same model across the mentioned datasets. Our approach, however, included a knowledge transfer process in each phase, where the λ between L_h and L_s was set to 0.01.