ALRMR-GEC: Adjusting Learning Rate Based on Memory Rate to Optimize the Edit Scorer for Grammatical Error Correction

Zhixiao Wu, Yao Lu*, Jie Wen*, Guangming Lu

Harbin Institute of Technology, Shenzhen, China
wzxnh24428@gmail.com, luyao2021@hit.edu.cn, jiewen_pr@126.com, luguangm@hit.edu.cn

Abstract

Edit-based approaches for **G**rammatical **E**rror **C**orrection (**GEC**) have attracted volume attention due to their outstanding explanations of the correction process and rapid inference. Through exploring the characteristics of the generalized and specific knowledge learning for GEC, we discover that efficiently training GEC systems with satisfactory generalization capacity prefers more generalized knowledge rather than specific knowledge. Current gradient-based methods for training GEC systems, however, usually prioritize minimizing training loss over generalization loss. This paper proposes the strategy of **A**djusting **L**earning **R**ate Based on **M**ermory **R**ate to optimize the edit-based **GEC** scorer (**ALRMR-GEC**). Specifically, we introduce the memory rate, a novel metric, to provide an explicit indicator for the model's state of learning generalized and specific knowledge, which can effectively guide the GEC system to adjust the learning rate timely. Extensive experiments, conducted by optimizing the published edit scorer on the BEA2019 dataset, have shown our ALRMR-GEC significantly enhances the model generalization ability with stable and satisfactory performance nearly irrespective of the initial learning rate selection. Also, our method can accelerate the training over tenfold faster in certain cases. Finally, the experiments indicate the memory rate introduced in our ALRMR-GEC guides the GEC editscorer to learn more generalized knowledge.

Introduction

Grammatical Error Correction (GEC) involves automatically identifying the errors and converting the source text to its clean version. Edit-based approaches have attracted volume attention due to their outstanding explanations of the correction process and rapid inference (Omelianchuk et al. 2020; Tarnavskyi, Chernodub, and Omelianchuk 2022). The editscorers (Sorokin 2022) use pre-trained Transformer-based models as encoders to identify the correctness of generated edits by GECToR (Omelianchuk et al. 2020) and achieve state-of-the-art quality. However, how to further unleash the potential of GEC models to get satisfactory performance remains an important issue.

GEC models aim to learn generalized knowledge like grammar rather than specific knowledge like word combinations and spelling. However, common gradients-based methods (Duchi, Hazan, and Singer 2011; Tieleman 2012; Kingma and Ba 2014) fail to guide models to enhance generalization ability considering **gradients alone do not guarantee the models to enhance generalization ability** since the gradients only provide the information about how to minimize the training loss. For example, the training period using Adamw optimizer with 1e-5 as the learning rate gets worse performance on the test dataset (Figure 1b) although it minimizes the training loss faster than the period using 1e-6 (Figure 1a). Therefore, researchers have to adjust the learning rate artificially to ensure satisfactory generalization performance, which is costly and labor-intensive.

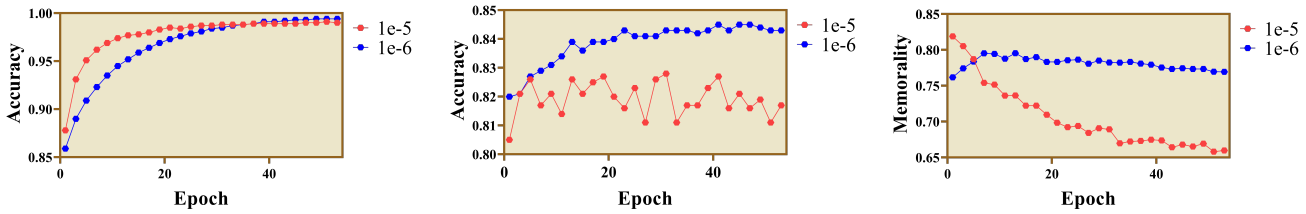
Finding a solution to separate generalized knowledge and specific knowledge learning is an urgent issue. In our analysis, generalized knowledge and specific knowledge represent the commonality and individuality of a model when learning from different data. For specific data, generalized knowledge can be acquired from other data while specific knowledge cannot. Therefore, we introduce the accuracy of the model on a specific subset not involved in later training (memorability) to indicate whether the current learning rate favors the generalized knowledge accumulation. Exploratory experiments are conducted in Figure 2 and the analysis can be seen as the following.

Firstly, we explore the influence of learning rates on memorability in Figure 2a and observe that memorability is notably diminished when the learning rate is far away from the favorable learning rate. *The above observation implies that memorability is a trade-off of two competing processes, designated as A and B. When the learning rate is excessively low (high), A (B) dominates and leads to low memorability.*

Secondly, we explore the characteristics of the above processes in Figure 2b. *Applied by small learning rates including 1e-7, process A dominates and the memorability tends to increase during the later training period. Correspondingly, applied by large learning rates including 5e-8, process B dominates and the memorability tends to decrease.*

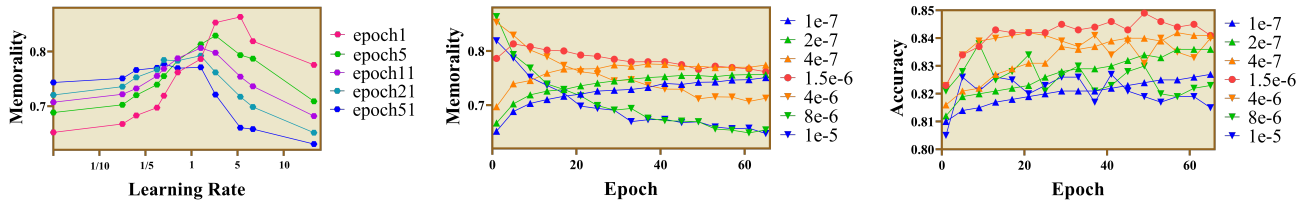
Finally, the correlation between generalization ability and learning rates can be seen in Figure 2c. Models trained with learning rates that keep memorability at a high level attain favorable performance on the test dataset. Specifically, 1.5e-6 is the favorable learning rate with nearly 84.8% accuracy on the test set and keeps memorability at a higher level

*corresponding author



(a) Accuracy on train dataset (b) Accuracy on test dataset (c) Trends of memorability

Figure 1: Gradients and memorability on GEC task.



(a) Learning rates and memorability (b) Trends of memorability (c) Accuracy on test dataset

Figure 2: The influence of learning rates on generalization ability and memorability. 1.5e-6 is named the **favorable learning rate** due to its favorable impact on the acquisition of generalized knowledge. The favorable learning rate is used as the benchmark ($x = 1$) and the value of the x-axis represents the multiples of the benchmark.

in Figure 2b compared to other learning rates. Upon comparative analysis of Figures 1 and 2, *memorability distinctively differentiates the performance of models across learning rates, surpassing the gradients as a more informative metric to guide the model to learn generalized knowledge.*

Based on the above observations, we realize the importance of memorability and its tight correlation to the favorable learning rate. We further explore the selection of benchmark data used to calculate the memorability (Figure 5) and solidify the memorability as ***Memory Rate, the accuracy of the particular subset selected from the train set that was classified incorrectly in the previous epoch but correctly in the current epoch and will not participate in the subsequent training periods***. Furthermore, ALRMR-GEC is proposed to optimize the editscorers automatically and the code can be seen at https://github.com/rearchwzx/ALRMR-GEC. The contributions of our work can be summarized by answering the following questions:

- Is there a metric for overcoming the drawback of gradients in enhancing the generalization ability?

  Yes, the proposed **memory rate provides an explicit indicator of the model's state of generalized knowledge learning**, which can guide the GEC systems to adjust the learning rates to the favorable learning rate timely and thus enhance the generalization ability of the editscorer.

- Can the learning rate be automatically adjusted efficiently and hence itself **no longer be a hyperparameter** that needs to be carefully adjusted artificially?

  Yes, the proposed ALRMR-GEC has two stages: the fast start stage and the slow adjustment stage. Based on the aforementioned stages, the ultimate performance is nearly regardless of the initial learning rate selection.

- Can the method further unleash the potential of pre-trained Transformer-based models for GEC?

  Yes, **we use ALRMR-GEC to further train more powerful and stable GEC editscorers effectively** based on the *Roberta-base* and the *Roberta-large* models. The optimized models can be stably maintained at the favorable generalization state. In specific cases, the time to reach a certain validation accuracy can be shortened to **1/10** compared to the origin training process.

Related Work

Pre-trained Transformer-based models, trained with massive language data, have demonstrated remarkable efficacy in NLP tasks (Hu et al. 2023; Zhong et al. 2022; Li et al. 2022, 2023a; Rossiello et al. 2023). Specifically, for GEC (Gong et al. 2022; Zhang et al. 2022; Fang et al. 2023), the architecture of multiple-layer multi-head attention mechanisms helps the systems (Gong et al. 2022; Sorokin 2022; Li et al. 2023b) exhibit enhanced feature extraction capabilities compared to their counterparts.

Most strategies used to unleash the potential of pre-trained Transformer-based models upon GEC are based on gradients (Smith and Topin 2019; Li and Arora 2020; Loshchilov and Hutter 2022; Iyer et al. 2023). Currently, Adam (Kingma and Ba 2014) and its variant AdamW (Loshchilov and Hutter 2017) are among the most popular optimizers (Schmidt, Schneider, and Hennig 2021). It retains the first and second moment information of parameters to fa-

cilitate adaptive learning step size. Shuaipeng Li (Li et al. 2024) establishes a scaling law between favorable learning rates and batch sizes for Adam-style optimizers. However, strictly following the derivative-guided route (Sutskever et al. 2013; Duchi, Hazan, and Singer 2011; Zeiler 2012; Tieleman 2012; Kingma and Ba 2014; Shazeer and Stern 2018; Smith and Topin 2019; Li and Arora 2020; Chen et al. 2024) does not guarantee the models to enhance generalization ability. We observe that some scientists (Fung, Yoon, and Beschastnikh 2018; Ozdayi, Kantarcioglu, and Gel 2021) adjust the learning rate based on global features to defend against backdoor attacks, which inspires us to find better metrics to adjust the learning rate.

Our Approach

In this section, a theoretical framework is constructed to elucidate the empirical phenomena and we introduce ALRMR-GEC, a novel strategy that dynamically adjusts the learning rate based on the characteristics of knowledge learning.

Specific Knowledge and Generalized Knowledge

The parameters of model can be denoted as $\theta_t = [\theta_t^1, \theta_t^2, \cdots, \theta_t^i, \cdots, \theta_t^n]$ and the knowledge learned at the t^{th} backpropagation can be expressed as $\Delta\theta_t$. Assuming an optimal target model is maximizing the generalization ability within the current architecture, parameters of the optimal model can be denoted as $\theta_{best} = [\theta_{best}^1, \theta_{best}^2, \cdots, \theta_{best}^i, \cdots, \theta_{best}^n]$.

Generalized knowledge refers to the knowledge that helps the current model to approach the optimal model. *Specific knowledge* refers to the knowledge that belongs to specific data that hinders the current model from approaching the optimal model. By introducing $[\alpha_t^1, \alpha_t^2, \cdots, \alpha_t^i, \cdots, \alpha_t^n](\alpha_t^i > 1)$ into definitions to better fit the real situation, the learned generalized and specific knowledge at the t^{th} backpropagation can be defined as follows:

$$\Delta\theta_t^{gen(spec)} = \begin{matrix}[\Delta\theta_{t,1}^{gen(spec)}, \Delta\theta_{t,2}^{gen(spec)}, \cdots, \\ \Delta\theta_{t,i}^{gen(spec)}, \cdots, \Delta\theta_{t,n}^{gen(spec)}]\end{matrix}, \tag{1}$$

$$\Delta\theta_{t,i}^{gen(spec)} = \begin{cases} \alpha_t^i * \Delta\theta_t^i & sgn(\Delta\theta_t^i) = (\neq) sgn(\Delta\theta_{t,i}^{best}) \\ (1-\alpha_t^i) * \Delta\theta_t^i & sgn(\Delta\theta_t^i) \neq (=) sgn(\Delta\theta_{t,i}^{best}). \end{cases} \tag{2}$$

$\Delta\theta_t^{gen}$ and $\Delta\theta_t^{spec}$ are complementary and together constitute the learned knowledge $\Delta\theta_t$. $sgn(\cdot)$ outputs -1 and 1 based on the direction of parameter variation. Distinct outputs ($sgn(\cdot) \neq sgn(\cdot)$) indicate the opposite directions.

Knowledge Learning

Symbols $SPEC$ and GEC represent the specific knowledge and generalized knowledge respectively. T represents the number of backpropagations during the subsequent training periods. The close-knit correlation between the edits instigates two adversarial periods is shown as follows:

Knowledge degradation : $\Delta X\_DE_{x,y}^Y$ in Eqns. 3 and 4 refers to the detrimental influence of knowledge Y learned at y^{th} backpropagation on knowledge X learned at x^{th} backpropagation ($x \leq y, X, Y \in \{SPEC, GEN\}$).

$$\Delta X\_DE_{x,y}^Y = \begin{matrix}[\Delta X\_DE_{x,y,1}^Y, \Delta X\_DE_{x,y,2}^Y, \cdots, \\ \Delta X\_DE_{x,y,i}^Y, \cdots, \Delta X\_DE_{x,y,n}^Y]\end{matrix} \tag{3}$$

$$\Delta X\_DE_{x,y,i}^Y = \begin{cases} \alpha_x^i * \Delta\theta_{y,i}^Y & sgn(\Delta\theta_{x,i}^X) \neq sgn(\Delta\theta_{y,i}^Y), \\ & sgn(\Delta\theta_{x,i}^X) = sgn(\Delta\theta_x^i) \\ (1-\alpha_x^i) * \Delta\theta_{y,i}^Y & sgn(\Delta\theta_{x,i}^X) \neq sgn(\Delta\theta_{y,i}^Y), \\ & sgn(\Delta\theta_{x,i}^X) \neq sgn(\Delta\theta_x^i) \\ 0 & others \end{cases} \tag{4}$$

Knowledge complement : $\Delta X\_CO_{x,y}^Y$ in Eqns. 5 and 6 refers to the positive influence of knowledge Ylearned y^{th} backpropagation on knowledge X learned at x^{th} backpropagation ($x \leq y, X, Y \in \{SPEC, GEN\}$).

$$\Delta X\_CO_{x,y}^Y = \begin{matrix}[\Delta X\_CO_{x,y,1}^Y, \Delta X\_CO_{x,y,2}^Y, \cdots, \\ \Delta X\_CO_{x,y,i}^Y, \cdots, \Delta X\_CO_{x,y,n}^Y]\end{matrix}, \tag{5}$$

$$\Delta X\_CO_{x,y,i}^Y = \begin{cases} \alpha_x^i * \Delta\theta_{y,i}^Y & sgn(\Delta\theta_{x,i}^X) = sgn(\Delta\theta_{y,i}^Y), \\ & sgn(\Delta\theta_{x,i}^X) = sgn(\Delta\theta_x^i) \\ (1-\alpha_x^i) * \Delta\theta_{y,i}^Y & sgn(\Delta\theta_{x,i}^X) = sgn(\Delta\theta_{y,i}^Y), \\ & sgn(\Delta\theta_{x,i}^X) \neq sgn(\Delta\theta_x^i) \\ 0 & others \end{cases} \tag{6}$$

As $\Delta X\_CO_{x,y}^Y$ and $\Delta X\_DE_{x,y}^Y$ have covered all variations of the parameters when $X = Y$ ($X, Y \in \{SPEC, GEN\}$) , it is unnecessary to consider the impact of $\Delta X\_CO_{x,y}^Y$ and $\Delta X\_DE_{x,y}^Y$ when $X \neq Y$. Treating knowledge independently renders the formulas clear and comprehensible.

Characteristic of specific knowledge learning : Specific knowledge is characterized by an average of zero on the gradients and randomness in directions. From the perspective of knowledge degradation, the detrimental influence of specific knowledge learned from subsequent data $\Delta SPEC\_DE_{i,t+1}^{SPEC}(t \geq i)$ continuously overwrites $\Delta\theta_i^{spec}$ until $\Delta\theta_i^{spec}$ is completely forgotten when $T \to \infty$, which can be depicted as follows:

$$\lim_{T\to\infty} \sum_{t=i}^{i+T} \Delta SPEC\_DE_{i,t+1}^{SPEC} = -\Delta\theta_i^{spec}. \tag{7}$$

Specific knowledge represents the non-commonality in parameter variation. From the perspective of knowledge complement, the positive influence of specific knowledge learned from subsequent data $\Delta SPEC\_CO_{i,t+1}^{SPEC}(t \geq i)$ is nearly non-existent when $T \to \infty$, which can not complement $\Delta\theta_i^{spec}$, resulting in a cumulative effect of 0. The above analysis can be depicted as follows:

$$\lim_{T\to\infty} \sum_{t=i}^{i+T} \Delta SPEC\_CO_{i,t+1}^{SPEC} = 0. \tag{8}$$

Characteristic of generalized knowledge learning : Generalized knowledge is characterized by a consistent directionality on the gradient, representing the commonalities. From the perspective of knowledge degradation, the detrimental influence of generalized knowledge learned from subsequent data $\Delta GEN\_DE_{i,t+1}^{GEN}(t \geq i)$ is nearly non-existent when $T \to \infty$ as the consistency in directions ensures that generalization knowledge does not cancel each other out, resulting in a cumulative effect of 0, which can be depicted as follows:

$$\lim_{T\to\infty} \sum_{t=i}^{i+T} \Delta GEN\_DE_{i,t+1}^{GEN} = 0. \tag{9}$$

Generalization knowledge represents the parameter variation that converge toward the optimal model. From the perspective of knowledge complement, the positive influence of generalized knowledge learned from subsequent data $\Delta GEN\_CO_{i,t+1}^{GEN}(t \geq i)$ continuously augments the unmastered generalization knowledge until θ_i is the same as the optimal model's parameters set θ_{best} when $T \to \infty$. The above analysis can be depicted as follows:

$$\lim_{T\to\infty} \sum_{t=i}^{i+T} \Delta GEN\_CO_{i,t+1}^{GEN} = \theta_{best} - \theta_i. \tag{10}$$

Memory Rate

$\Delta SPEC_i^{epoch\_i}$ represents the accumulation of specific knowledge learned at i^{th} backpropagation during the current epoch. Specifically, $epoch\_i$ refers to the number of aftermost backpropagation during the epoch which includes i^{th} backpropagation. The detrimental and positive influence of specific knowledge learned from subsequent data can be represented as $\Delta SPEC\_DE_{i,t+1}^{SPEC}$ and $\Delta SPEC\_CO_{i,t+1}^{SPEC}(t \geq i)$, respectively. The mathematical description of the above process can be presented as follows:

$$\begin{aligned}\Delta SPEC_i^{epoch\_i} = \Delta\theta_i^{spec} + \sum_{t=i}^{epoch\_i} (\Delta SPEC\_DE_{i,t+1}^{SPEC} \\ + \Delta SPEC\_CO_{i,t+1}^{SPEC}).\end{aligned} \tag{11}$$

A large amount of data is one of the characteristics of the training set. Therefore, for the subset trained at the early stage (represented as A), the number of backpropagations during the subsequent training periods T can be seen as $T \to \infty$. According to Eqns. 7 and 8, its specific knowledge is almost overwritten and cannot be accumulated by the latter learned specific knowledge, resulting in a cumulative effect close to 0, which can be depicted as follows:

$$\Delta SPEC_i^{epoch\_i} \approx 0\ (i \in A). \tag{12}$$

However, the size of the dataset is limited. Therefore, for the fresh subset trained at $\{i|i \notin A\}$, the number of backpropagations during the subsequent training periods T can not be seen as $T \to \infty$. Therefore, its specific knowledge can not be completely overwritten by the latter learned specific knowledge, which contributes to the principal part of the learned specific knowledge at the current epoch, resulting in a non-zero cumulative effect. The mathematical description can be presented as follows:

$$\Delta SPEC_i^{epoch\_i} > 0\ (i \notin A). \tag{13}$$

$\Delta GEN_i^{epoch\_i}$ represents the accumulation of generalized knowledge learned at i^{th} backpropagation during the current epoch. The detrimental and positive influence of generalized knowledge learned from subsequent data can be represented as $\Delta GEN\_DE_{i,t+1}^{GEN}$ and $\Delta GEN\_CO_{i,t+1}^{GEN}(t \geq i)$, respectively. The mathematical description of the above process can be presented as follows:

$$\begin{aligned}\Delta GEN_i^{epoch\_i} = \Delta\theta_i^{gen} + \sum_{t=i}^{epoch\_i} (\Delta GEN\_DE_{i,t+1}^{GEN} \\ + \Delta GEN\_CO_{i,t+1}^{GEN}).\end{aligned} \tag{14}$$

The performance of models upon the data trained at i^{th} backpropagation after the current epoch is a trade-off adversarial process based on $\Delta GEN_i^{epoch\_i}$ and $\Delta SPEC_i^{epoch\_i}$. $\Delta GEN_i^{epoch\_i}$ contributes to the optimization of the model with the hindrance of $\Delta SPEC_i^{epoch\_i}$ $(i \notin A)$.

$$\Delta\theta_i^{epoch\_i} = \Delta GEN_i^{epoch\_i} + \Delta SPEC_i^{epoch\_i}. \tag{15}$$

Based on improper small learning rates, the unsatisfactory generalized knowledge $\Delta\theta_i^{gen}$ learned at i^{th} backpropagation leads to a low memory rate. The memory rate will increase as the insufficient generalized knowledge will be supplemented by other data in subsequent training periods according to Eqn. 14. Based on improper large learning rates, the preference of models on fresh data is strengthed. The accumulation of specific knowledge, depicted in Eqn. 13, prevails compared to the accumulation of generalized knowledge. The memory rate is expected to decrease in the latter training periods according to Eqn.12.

Learning Rate Evolution Strategy

At the fast-start stage, ALRMR-GEC introduces a backtracking mechanism to initiate improper training and find appropriate learning rates with a complexity of $O(log_2 n)$. At the fine-turning stage, ALRMR-GEC keeps the learning rate at a favorable status to elaborately optimize the model and ensure its stability, thereby preventing overfitting.

Algorithm 1 ALRMR-GEC

Require : Learning rate lr>0, the editscorer $model_\theta$
Require : $modelstage$ differentiates the fast-start phase 0 and the fine-tune phase 1
Require : $correctflag$ differentiates the recalculation time 0 and the adjustment time 1
Initialize : $correctflag$, $modelstage$ = 0
while training **do**
  if $correctflag$ is 0 and $modelstage$ is 0 **then**
    Backtrack $model_\theta$ to the initial state.
  end if
  if $correctflag$ is 1 **then**
    Train $model_\theta$ with lr
    Calculate the memory rate $nowmero$
    if $modelstage$ is 0 **then**
      if The variation or the continuous change of $nowmero$ exceeds the threshold **then**
        Use binary search to find the favorable learning rate
        $correctflag$ = 0
      else
        $correctflag$, $modelstage$ = 1
      end if
    else
      if The variation or the continuous change of $nowmero$ exceeds the threshold **then**
        lr = lr*$\alpha(\beta)$, α and β (<1)
        $correctflag$ = 0
      else
        $correctflag$ = 1
      end if
    end if
  else
    Update the benchmark dataset
    Calculate the memory rate $lastmero$
    $correctflag$ = 1, $stagenum$ = $lastmero$
  end if
end while

Model		*Roberta-base*		*Roberta-large*	
		Pie_bea-gector	ALRMR-GEC	Clang_large_ft2	ALRMR-GEC
Threshold = 0.5	P	57.43	**65.45**	63.18	**65.95**
	R	31.68	**32.31**	**37.40**	35.27
	F	49.40	**54.31**	55.53	**56.18**
	Acc(%)	78.51	**85.11**	84.80	**85.90**
Threshold = 0.7	P	63.24	**65.92**	66.60	**66.67**
	R	28.16	**31.76**	**35.67**	34.80
	F	50.63	**54.25**	**56.76**	56.35
	Acc(%)	82.62	**85.16**	86.01	**86.13**
Threshold = 0.9	P	**76.20**	67.12	**72.17**	67.67
	R	13.97	**31.16**	30.52	**33.84**
	F	40.29	**54.53**	**56.69**	56.40
	Acc(%)	83.24	**85.33**	**86.59**	86.33

Table 1: Performance of ALRMR-GEC based on *Roberta-base* and *Roberta-large* using *Faster Simultaneous Decoding*.

Model		Threshold = 0.7		Threshold = 0.8		Threshold = 0.9	
		origin	ALRMR-GEC	origin	ALRMR-GEC	origin	ALRMR-GEC
Stage = 1	P	**72.25**	71.28	**75.14**	73.33	**76.90**	75.68
	R	17.70	**19.90**	16.19	**19.52**	13.95	**18.63**
	F	44.70	**47.01**	43.48	**47.27**	40.43	**46.94**
Stage = 3	P	**68.41**	67.49	**72.34**	70.12	**74.99**	73.56
	R	27.80	**31.80**	24.78	**30.84**	20.31	**28.76**
	F	52.94	**55.12**	52.27	**55.89**	48.74	**56.09**
Stage = 5	P	**67.72**	66.57	**71.87**	69.42	**74.93**	73.08
	R	30.24	**34.41**	26.68	**33.23**	21.62	**30.84**
	F	54.27	**56.08**	53.68	**57.00**	50.18	**57.37**
Stage = 7	P	**67.46**	66.59	**71.69**	69.37	**74.89**	73.11
	R	30.99	**35.26**	27.24	**33.98**	22.00	**31.53**
	F	54.60	**56.54**	54.05	**57.41**	50.57	**57.85**

Table 2: Performance of ALRMR-GEC based on *Roberta-base* using *Better Stagewise Decoding*.

Experimental Results and Analysis

Preliminary

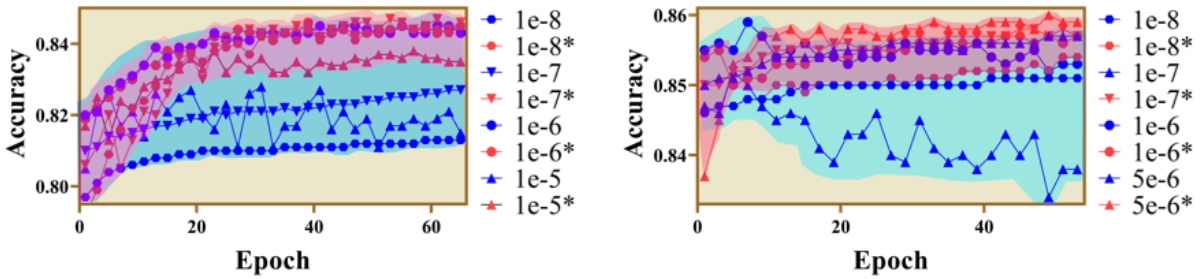
The implementation follows the proposed editscorer and employs the same training data and experimental setup (Sorokin 2022). The gector_variants are generated using edit generators on the BEA 2019 Shared Task data. The Base strategy in Figure 5 is used as the default approach to select the benchmark. Furthermore, there are two decoding strategies used in our experiments. *Faster Simultaneous Decoding* is an offline approach in which the editscorer calculates a collection that satisfies the criteria whose probability exceeds the predefined threshold and then chooses the edits that perform higher scores and do not contradict other edits. At *Better Stagewise Decoding*, the editscorer selects the most probable edit at first. Then the editscorer applies it to the current input sentence and removes all the edits with intersecting spans repeatedly until the most probable edit is "do nothing" or its probability is below the threshold.

Superiority of Our Approach

Superiority of Generalization Ability: The models optimized by ALRMR-GEC can surpass original editscorers within *Roberta-base* models both in *Faster Simultaneous Decoding* (Table 1) and *Better Stagewise Decoding* (Table 2). Our approach improves Acc by 6.9% when using 0.5 as the threshold during the inference period in *Faster Simultaneous Decoding*. In *Better Simultaneous Decoding*, the proposed ALRMR-GEC enhances the Roberta-base models in recall and f-measure with a minor decrease in precision. The above observations suggest that the memory rate can enhance the generalization ability of GEC models.

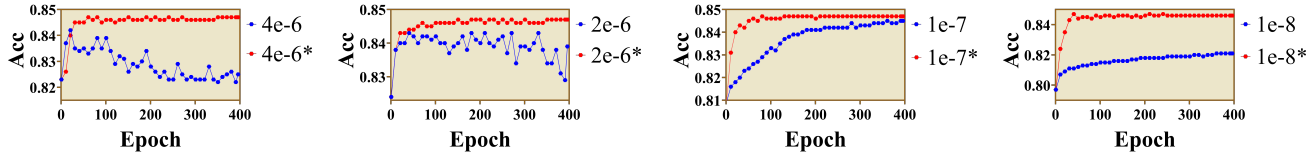
The models optimized by ALRMR-GEC can not get the same amazing improvement within *Roberta-large* models compared to *Roberta-base* models in *Faster Simultaneous Decoding*. We speculate that this is because smaller models are more reliant on the choice of learning rate and the memory rate is based on the accuracy so that the models are more inclined to improve the Acc rather than the other metrics. Adjusting the definition of the memory rate may alter the direction of model improvement.

Superiority of Efficiency: The red regions exhibit two characteristics compared to the blue regions: a steady increase and a narrower range (Figure 3), which respectively align with our method's advantages in stability and efficiency. Therefore, the editscorers can be optimized nearly regardless of the initial learning rate selection and do not need to be adjusted manually. Specifically, ALRMR-GEC



(a) Roberta-base (b) Roberta-large

Figure 3: Performance with different selections of initial learning rates. Accuracy represents the accuracy of the trained model upon the test dataset. * represents the training period using the ALRMR-GEC method. **Blue** represents the area of the origin training period and **Red** depicts the area of the training period using ALRMR-GEC.



(a) Improvement upon overfitting (b) Improvement upon underfitting

Figure 4: Performance based on *Roberta-base* model with different selections of initial learning rates during 400 epochs. Acc represents the accuracy of the trained model upon the test dataset. * represents the training period using the ALRMR-GEC method. In the actual training environment, each training epoch approximately consumes 25 minutes.

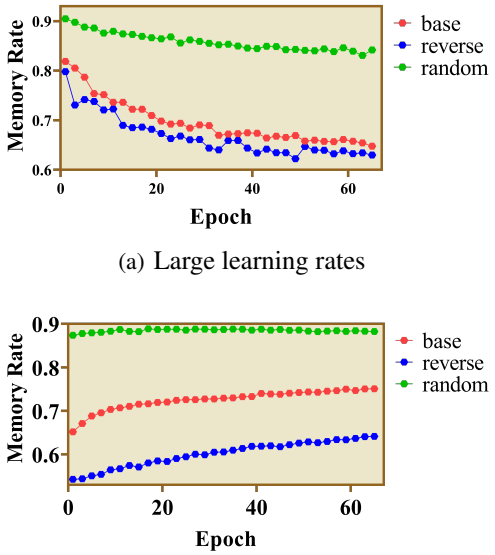
can accelerate the training period especially when the learning rate is improperly small. Figure 4b shows that even after 400 epochs (nearly 166 hours) of training, the model trained by 1e-8 is still far from achieving a satisfactory performance. However, it takes only 37 epochs (nearly 15 hours) to train the *Roberta-base* models to touch 84% on validation accuracy when optimized by ALRMR-GEC. The results indicate that the proposed ALRMR-GEC can accelerate the models to learn the generalized knowledge.

Superiority of Stability: According to Figure 4, the related models suffer from overfitting in the origin training period even if both 2e-6 and 4e-6 are favorable learning rates at the beginning. Optimized by ALRMR-GEC, *Roberta-base* models can be stably maintained at the peak performance state upon the validation accuracy (84.8%). Therefore, overfitting is not an inevitable trend in model training. Properly adjusting the learning rate enables the model to increase accuracy on the training set without compromising its accuracy on the test set. Overfitting arises owing to the acquisition of detrimental specific knowledge. The results reply that the memory rate constrains the model from overly focusing on specific knowledge within GEC sentences.

Ablation Study

1. The approaches of choosing the benchmark dataset

According to Figure 5a, for the Base/Reverse strategy, the memory rate varies from 0.6 to 0.8 whose fluctuations become much more pronounced compared to the Random



(a) Large learning rates

(b) Small learning rates

Figure 5: The approaches of choosing the benchmark dataset to calculate the memory rate. **Base** (**Reverse**) refers to selecting the subset of the training set that was classifed incorrectly (correctly) in the previous epoch but correctly (incorrectly) in the current epoch. **Random** refers to randomly selecting a subset of the training set.

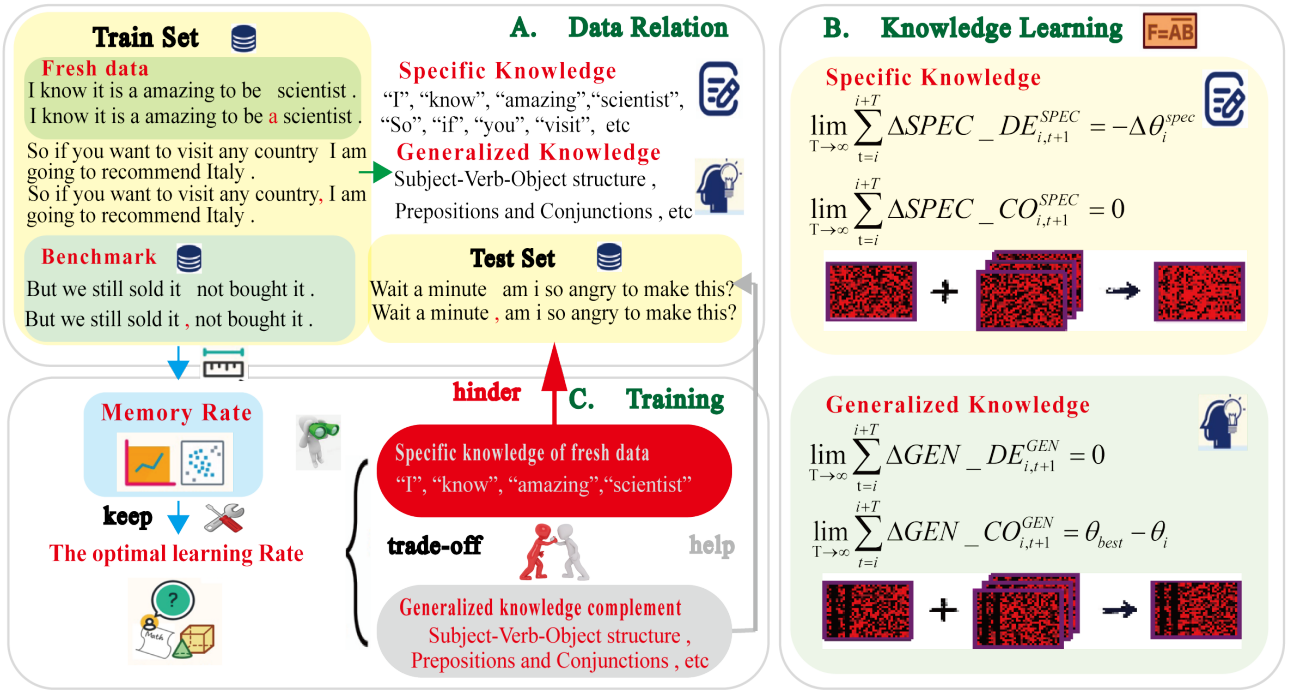


Figure 6: Visualization of knowledge learning. $\Delta\theta^i = \Delta\theta_t^i - \Delta\theta_0^i > 0$ is named as the positive direction of θ^i. Therefore, the parameter variation in the negative (positive) direction is reflected by **black** (**red**). **Data Relation** investigates the source of learned knowledge. Secondly, **Knowledge Learning** experimentally demonstrates the characteristics of specific and generalized knowledge accumulation. Finally, the entire training process and the analysis of memory rate are reviewed in **Training**.

strategy in which the memory rate varies from 0.85 to 0.9. According to Figure 5b, the memory rate based on Random keeps 0.9 during the training and fails to reflect the state of generalized knowledge learning. Conversely, the benchmark datasets on the Base/Reverse strategy contain more knowledge that has not been grasped by the model and the variation in accuracy can still intuitively reflect the suitability of the hyperparameters adopted in the current training period.

2. Visualization of knowledge learning

Data Relation: For Benchmark in Figure 6A, inserting "," into the "sold it" and "not bought it." needs the model to learn the generalized knowledge about the Subject-Verb-Object structure. The specific knowledge of Fresh data ("I", "know", etc) contributes to the major part of the learned specific knowledge during the current training period (Eqn. 13). The specific knowledge learned at the early training ("So", "if", etc) will be forgotten according to Eqn. 8. But the generalized knowledge can be accumulated with proper learning rates according to the discussion about Eqn. 15.

Knowledge Learning: Within large learning rates, the model's preference for fresh data dominates, leading to significant parameter variation that mainly reflects the characteristics of specific knowledge in different fresh data. According to Figure 6B, the directions of parameter variation in different specific knowledge are so random that their cumulative effects cancel each other out. Conversely, with small learning rates, the model's emphasis on accumulating generalized knowledge prevails. The similar parameters variation implies the universality of generalized knowledge.

Training: On one hand, higher learning rates accelerate the accumulation of generalized knowledge. On the other hand, the bias of the specific knowledge ("I", "know", "amazing", etc) is also intensified as the gradients of fresh data focus on details about word spelling without considering the learned grammar (Subject-Verb-Object structure) used in the sentences "But we still sold it, not bought it". The favorable learning rate is a trade-off point of the above adversarial periods, which can be depicted in Figure 6C. Memory rate, analogous to the performance of learners upon error-correction notebooks without reviewing in human grammar learning, serves as a global feature to guide the ALRMR-GEC to indicate whether the current learning rate favors the generalized knowledge accumulation.

Conclusion

Grammatical Error Correction aims to learn generalized knowledge like grammar rather than specific knowledge like word combinations and spelling. Generalized knowledge and specific knowledge represent the commonality and individuality of a model when learning from different data. Generalized knowledge can be acquired from other data while specific knowledge cannot. Based on this analysis, we introduce memory rate to guide the ALRMR-GEC to find the optimal learning rate timely.

Acknowledgments

This work was supported in part by the NSFC fund (NO. 62206073, 62176077), in part by the Shenzhen Key Technical Project (NO. JSGG20220831092805009, JSGG20220831105603006, JSGG20201103153802006, KJZD20230923115117033, KJZD20240903100712017), in part by the Guangdong International Science and Technology Cooperation Project (NO. 2023A0505050108), in part by the Shenzhen Fundamental Research Fund (NO. JCYJ20210324132210025), and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (NO. 2022B1212010005), and in part by the Natural Science Foundation of Shenzhen General Project under Grant JCYJ20240813110007010, in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515010893, in part by the Shenzhen Doctoral Initiation Technology Plan under Grant RCBS20221008093222010, in part by the Shenzhen Pengcheng Peacock Startup Fund.

References

Chen, X.; Liang, C.; Huang, D.; Real, E.; Wang, K.; Pham, H.; Dong, X.; Luong, T.; Hsieh, C.-J.; Lu, Y.; et al. 2024. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Fang, T.; Liu, X.; Wong, D. F.; Zhan, R.; Ding, L.; Chao, L. S.; Tao, D.; and Zhang, M. 2023. Transgec: Improving grammatical error correction with translationese. In *Findings of the Association for Computational Linguistics: ACL 2023*, 3614–3633.

Fung, C.; Yoon, C. J.; and Beschastnikh, I. 2018. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*.

Gong, P.; Liu, X.; Huang, H.; and Zhang, M. 2022. Revisiting grammatical error correction evaluation and beyond. *arXiv preprint arXiv:2211.01635*.

Hu, J.; Guo, D.; Liu, Y.; Li, Z.; Chen, Z.; Wan, X.; and Chang, T.-H. 2023. A simple yet effective subsequence-enhanced approach for cross-domain NER. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12890–12898.

Iyer, N.; Thejas, V.; Kwatra, N.; Ramjee, R.; and Sivathanu, M. 2023. Wide-minima density hypothesis and the explore-exploit learning rate schedule. *Journal of Machine Learning Research*, 24(65): 1–37.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, B.; Du, Q.; Zhou, T.; Jing, Y.; Zhou, S.; Zeng, X.; Xiao, T.; Zhu, J.; Liu, X.; and Zhang, M. 2022. ODE transformer: An ordinary differential equation-inspired model for sequence generation. *arXiv preprint arXiv:2203.09176*.

Li, B.; Yu, D.; Ye, W.; Zhang, J.; and Zhang, S. 2023a. Sequence generation with label augmentation for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13043–13050.

Li, S.; Zhao, P.; Zhang, H.; Sun, X.; Wu, H.; Jiao, D.; Wang, W.; Liu, C.; Fang, Z.; Xue, J.; et al. 2024. Surge Phenomenon in Optimal Learning Rate and Batch Size Scaling. *arXiv preprint arXiv:2405.14578*.

Li, Y.; Liu, X.; Wang, S.; Gong, P.; Wong, D. F.; Gao, Y.; Huang, H.-Y.; and Zhang, M. 2023b. TemplateGEC: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6878–6892.

Li, Z.; and Arora, S. 2020. AN EXPONENTIAL LEARNING RATE SCHEDULE FOR DEEP LEARNING. In *8th International Conference on Learning Representations, ICLR 2020*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Loshchilov, I.; and Hutter, F. 2022. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*.

Omelianchuk, K.; Atrasevych, V.; Chernodub, A.; and Skurzhanskyi, O. 2020. GECToR–Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 163–170.

Ozdayi, M. S.; Kantarcioglu, M.; and Gel, Y. R. 2021. Defending against backdoors in federated learning with robust learning rate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9268–9276.

Rossiello, G.; Chowdhury, M. F. M.; Mihindukulasooriya, N.; Cornec, O.; and Gliozzo, A. M. 2023. Knowgl: Knowledge generation and linking from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 16476–16478.

Schmidt, R. M.; Schneider, F.; and Hennig, P. 2021. Descending through a crowded valley-benchmarking deep learning optimizers. In *International Conference on Machine Learning*, 9367–9376. PMLR.

Shazeer, N.; and Stern, M. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, 4596–4604. PMLR.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 369–386. SPIE.

Sorokin, A. 2022. Improved grammatical error correction by ranking elementary edits. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11416–11429.

Sutskever, I.; Martens, J.; Dahl, G.; and Hinton, G. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, 1139–1147. PMLR.

Tarnavskyi, M.; Chernodub, A.; and Omelianchuk, K. 2022. Ensembling and Knowledge Distilling of Large Sequence Taggers for Grammatical Error Correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3842–3852.

Tieleman, T. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2): 26.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, Y.; Zhang, B.; Li, Z.; Bao, Z.; Li, C.; and Zhang, M. 2022. SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. *arXiv preprint arXiv:2210.12484*.

Zhong, M.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11765–11773.