L1-aware Grammatical Error Correction via Multitasking with Native Language Estimation

Yuehao Yuan
The University of Tokyo
yuanyuehao@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

1 Introduction

Grammatical errors of second-language learners depend strongly on their native languages (L1) [1]. For example in learning English as a second language, Finnish-L1 learners tend to overgeneralize the usage of the preposition *in* [2], while Chinese-L1 learners tend to make errors on verb tense and form, possibly because Chinese lacks verb inflection [3]. In performing gramamtical error correction (GEC), it is therefore important to take the writer' L1 for a given input into consideration.

Several researchers have proposed L1-aware GEC methods [2, 4]. Essentially, these methods regard text written by specific L1 writers as independent domains, and perform domain adaptation such as fine tuning for neural GEC models [4]. Because this approach assumes that the writers' L1s are known, it cannot be directly applied to text whose writers' L1 are unknown (e.g., Web text written by anonymous second-language learners) and text whose writers' L1s are not covered by the training data of the L1-aware GEC model.

To address the above limitations of the existing L1-aware GEC models, we propose an L1-aware GEC model that simultaneously estimates the writer' L1 (and the language family[1)] to which it belongs to) in reading (encoding) input to be corrected, in order to make the encoder be aware of the writer's L1 (Figure 1). In contrast to the existing L1-aware GEC models that explicitly takes advantage of the writers' L1s, our method demands L1 information only in training the GEC model, and it can be therefore applied to text whose writer's L1 is unknown.

We evaluate the effectiveness of our GEC model on Lang-8 dataset [5]. Experimental results confirmed that our method can perform accurate L1-aware GEC without assuming L1 information in evaluation.

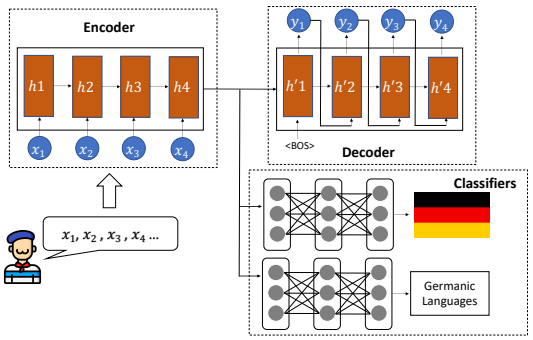

Figure 1 Our L1-aware GEC model with multitasking with L1 estimation.

2 Related work

Inspired by the observation that the L1s of second-language learners of English strongly influence the expression and language usage of English [1], researchers set up various NLP tasks such as native language identification [6] and linguistic typology prediction [7], and also exploited the observation in solving our target task, grammatical error correction (GEC) [2, 4].

Since the distributions of grammatical errors depend on the learners' L1s [8], L1-aware GEC has been studied as a domain adaptation problem. Shamil purposed neural network joint models, and used L1-specific data for fine-tuning [2]. Maria and Joel considered L1 and proficiency at the same time and demonstrate the effective personalized GEC [4]. Although these approaches successfully improved the GEC performance, they are not applicable to text written by unknown-L1 writers.

We have two public datasets for training L1-aware GEC models. Lang-8 [5] dataset is extracted from a social networking site that aims to help users learn each other's L1.

1) Precisely speaking, in this study, we also utilize subgroups of common language families, although we consistently use the term language family.

L1	Lang-8	FCE
Japanese	516,240	1337
Korean	65,249	1331
Chinese	61,682	1331
Russian	14,640	1500
Spanish	5688	2897
Polish	3075	1450
German	2160	1207
French	1978	2374
Dutch	0	47

Table 1 The number of sentences written by specific-L1 learners of English in two GEC datasets (after preprocessing).

CLC-FCE [9] compiles exam scripts in English written by students from various countries. These datasets, however, have some shortcomings for L1-aware GEC; the lack and imbalance of L1-specific data, as shown in Table 1. Although the fine-tuning used in neural L1-aware GEC models alleviates the data sparseness problem to a certain extent, it will not be effective when few or no annotated data is available as for Dutch.

3 Proposal

A straightforward way to perform L1-aware GEC on text whose writer's L1 is unknown is to first perform native language identification on the text and apply the GEC model trained for the estimated L1. However, this pipeline approach will be affected by error propagation, since accurate natural language identification is unrealistic when the input length is small. Even if input length is long, the input can be written by more than a single writers of various L1; for example, the paper you are reading now has been collaboratively written by Chinese and Japanese second-language learners of English.

In this study, we therefore propose to implicitly consider L1 of the writer of the input, and train a neural seq2seq-based GEC model [10] using a multitask learning with native language (L1) estimation as the auxiliary task. Specifically, the encoder of the GEC model is abused to guess the L1 of the writer of input text so that it can encode text while being aware of the writer's L1. In testing, the obtained L1-aware encoder can be applied to any text since it does not require L1 annotation to the test input. In what follows, we explain the auxiliary tasks on L1 estimation (§ 3.1), the model structure (§ 3.2), and the training procedure (§ 3.3).

3.1 Auxiliary tasks

Keeping in mind that native language identification from short input is technically difficult, we set two L1 estimation tasks as auxiliary tasks: one task identifies the writer's L1 for a given input, and the other task identifies the language family of the writer's L1. By using the language families with coarse-grained labels instead of languages themselves, we will be able to make the training on the extreme classification of natural language identification more stable [11].

As far as we know, no one tried to use language family of L1 in L1-aware GEC. We therefore performed preliminary experiments to investigate the effectiveness of the language family in L1-aware GEC in the exising L1-aware GEC. Using the Lang-8 datasets, we trained a neural seq2seq-based GEC model [10] by fine-tuning a GEC model trained from the Japanese-L1 data using various L1 data and evaluated the performance of the resulting GEC models on the Spanish-L1 data. The results of this preliminary experiment revealed that GEC data written by learners whose L1s are in the same language family as Spanish is also effective in fine tuning.

3.2 Model structure for multitasking

The key issue in a performing multitask learning is what to share in the model structure to solve the main and auxiliary tasks. In this study, we adopt a neural seq2seq-based GEC model [10, 12] as the basic model structure. To perform a multitask learning with native language estimation, we share the encoder of the main GEC task with the classifiers of L1 (and its language family) estimation, and add a feedforward neural network to perform the classification on the top of the GEC encoder. The overview of our GEC model can be seen back in Figure 1, which illustrates the the following computation process:

$$h_t^{(enc)} = \text{encoder}(x_1^t) \quad (1)$$

$$h_{t'}^{(dec)} = \text{decoder}(y_{t'-1}, h_{t'-1}^{(dec)}; \boldsymbol{h}^{enc}) \quad (2)$$

$$P(y_{t'}) = \text{softmax}(h_{t'}^{(dec)}) \quad (3)$$

$$P(c) = \text{softmax}(\text{FFNN}(h_n^{(enc)})) \quad (4)$$

We denote the vectors from the hidden state of the topmost layer of encoder as $h_t^{(enc)}$. Eq.1 through Eq.3 shows the process of generating output $\boldsymbol{y}$ (corrected input, the main

	Sino-Tibetan	Macro-Altaic		Slavic		Romance		
Dataset	Chinese	Japanese	Korean	Russian	Polish	Spanish	French	Italian
Training	60,018	30,007	30,015	10,013	2,002	4,002	1,001	1,000
Test	12,004	6,011	6,005	2,002	400	803	402	401

Table 2 Statistics of each L1-specific data in training dataset and test dataset; Chinese-, Japanese- and Korean-L1 data are sampled to perform experiments effectively.

	Sino-Tibetan	Macro-Altaic		Slavic		Romance		
Models	Chinese	Japanese	Korean	Russian	Polish	Spanish	French	Italian
Baseline	23.14	21.94	25.40	**28.83**	**33.51**	**22.91**	21.46	**24.64**
MTL (only L1)	23.12	21.53	26.12	27.92	31.97	22.06	23.04	22.95
MTL (only LF)	23.07	22.04	26.08	28.12	33.04	22.24	22.81	23.67
MTL (both L1 and LF)	**23.25**	**22.35**	**26.60**	28.56	33.42	22.30	**23.27**	23.19

Table 3 $F_{0.5}$ comparison of models on each L1, FT denote the fine-tuning model corresponding to the selected L1.

task), Eq.4 uses the encoer's output to predict the class (L1 or its language family). When we train these tasks simultaneously, the training of every task will contribute to updating parameters of the GEC encoder; in other words, the model reads (encodes) text while being aware of the writer's L1 and generates (decodes) corrected input from the L1-aware encoding of the input.

3.3 Training procedure

In the multitask learning, it is important to control the impact of auxiliary tasks on updating the model's parameters. If we keep to incorporate the loss obtained from the auxiliary tasks during whole training, it may harm the performance of the main task. Therefore, we adopt task-wise early stopping [13], a strategy to stop training of auxiliary tasks prior to the main task to suppress the extra impact of auxiliary tasks.

4 Experiments

To confirm the effectiveness of our GEC model via multitask learning with native language estimation, we utilize the Lang-8 dataset [5] to compare our model with the general seq2seq-based GEC model that does not perform the multitasking. We evaluate the GEC models using the Precision, Recall, and $F_{0.5}$ measure, computed by the M2 scorer [14].

4.1 Settings

In this experiment, we use the Lang-8 dataset [5] that consists of English text written by Chinese, Japanese, Korean, Spanish, Russian, Polish, French, and Italian-L1 learners of English. Since the the size of L1 data for Japanese, Chinese, and Korean is much more than the other languages (Table 1), we randomly sampled part of data for those resource-rich languages for efficient experimentation. We summarized the statistics of the reduced dataset in Table 2.

To compare with the baseline model, we prepare three variants of multitask-learning models using the two auxiliary tasks; namely, the one with L1 identification, the one with L1's language family identification, and the one with L1 and its language family identification. The baseline is the neural seq2seq GEC model [10] that is identical to our model without multitasking.

The encoder and the decoder of the GEC model are three-layer bi-directional LSTMs with 200-dimensional hidden states for each layer. The feed-forward networks for the auxiliary tasks have a 200-dimensional input layer and a 256-dimensional hidden layer.

To find the best model by the task-wise early stopping [13], we set up multiple stop points from 5 to 30 with interval of 5. When the training reaches the epoch corresponding to the stop points, we fork the training process to start the training with the GEC loss only. We choose the best stop points that maximized the GEC performance on each L1 dev data.

4.2 Results

Table 3 lists performance of the GEC models on each L1-specific data. We can see from this result that our method

could successfully improved L1s with more training data. However, we could not improve the GEC performance for L1s with less training data except French. This is possibly because the native language estimation module in our model tends to estimate dominant L1s from input, which affects a negative impact on the GEC performance of European languages.

On the other hand, by comparing the performance of MTL (only L1) and MTL (only LF) on Japanese-L1 and Polish-L1 data, we can find that performance of the latter model has improved significantly, which implies that the introduction of language family is helpful for the GEC model on L1s which is more difficult to distinguish due to less training data.

Finally, we investigated the performance of two classification tasks on the test data. We notice that the accuracy of the L1 identification task is 49.90%, while the accuracy of the L1 language family identification is 57.85%. The majority-class baselines for the two identification tasks is 43.47%. Therefore, the accuracy of the classifiers is not very high, especially for the L1 classifier. Although the accuracy of native language estimation tasks is low, considering those tasks as auxiliary tasks have greatly contributed to solving grammatical error correction.

5 Conclusion and future work

In this study, we propose an L1-aware GEC model via multitask learning with native language estimation. Our model implicitly uses L1s in contrast to the existing methods that is based domain adaptation and explicitly uses L1s in evaluation. To mitigate unstable training caused by the imbalance L1 data on GEC, we consider two tasks on native language estimation; namely, native language (L1) identification and L1's language family identification. Experimental results confirmed that our method achieves improvement over the baseline.

References

[1]Michael Swan and Bernard Smith. *Learner English: A Teacher's Guide to Interference and Other Problems*. Cambridge Handbooks for Language Teachers. Cambridge University Press, 2 edition, 2001.

[2]Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1901–1911, Austin, Texas, November 2016.

[3]Alison Bernstein. *The School Review*, Vol. 86, No. 2, pp. 292–294, 1978.

[4]Maria Nadejde and Joel Tetreault. Personalizing grammatical error correction: Adaptation to proficiency level and L1. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 27–33, Hong Kong, China, November 2019.

[5]Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 147–155, Chiang Mai, Thailand, November 2011.

[6]SEAN MASSUNG and CHENGXIANG ZHAI. Non-native text analysis: A survey. *Natural Language Engineering*, Vol. 22, No. 2, p. 163–186, 2016.

[7]Yevgeni Berzak, Roi Reichart, and Boris Katz. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 21–29, Ann Arbor, Michigan, June 2014.

[8]Yevgeni Berzak, Roi Reichart, and Boris Katz. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 94–102, Beijing, China, July 2015.

[9]Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 180–189, Portland, Oregon, USA, June 2011.

[10]Tao Ge, Xingxing Zhang, Furu Wei, and Ming Zhou. Automatic grammatical error correction for sequence-to-sequence text generation: An empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6059–6064, Florence, Italy, July 2019.

[11]Daisuke Oba, Naoki Yoshinaga, Shoetsu Sato, Satoshi Akasaki, and Masashi Toyoda. Modeling personal biases in language use by inducing personalized word embeddings. 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)*, pp. 2102–2108, Minneapolis, Minnesota, June 2019.

[12]Tao Ge, Furu Wei, and Ming Zhou. Reaching human-level performance in automatic grammatical error correction: An empirical study. *CoRR*, Vol. abs/1807.01270, , 2018.

[13]Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pp. 94–108, Cham, 2014. Springer International Publishing.

[14]Daniel Dahlmeier and Hwee Tou Ng. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 568–572, Montréal, Canada, June 2012.