Fine-tuning GEC Model Based on Language Family Corpus

Yitao Liu*
Zhejiang Guangshan Vocational and Technical University of Construction, Zhejiang, China
yitao.liu@zjgsdx.edu.cn

Mark Dras
Macquarie University
Sydney, Australia
mark.dras@mq.edu.au

Abstract

It is widely known that the first language (L1) of the English learners will influence their language study, causing them make to biased errors. However, it is relatively limited for the research of using the L1 information to improve Grammatical Error Correction (GEC) models. Among the limited research, a common method is to train a set of GEC models, and each model is trained by a corpus from one (and only one) specific L1 background. This method has been proven efficient, while the waste of the training / fine-tuning data makes it suffer from the data limitation issue. This paper introduces a novel method to address this issue by exploiting the linguistic similarities between a language family and its member languages. We expand the fine-tuning data from one specific L1 background to its language family one, making the quantity increase exponentially. We use the Italic language family corpus as our language family corpus and experiment with two approaches facing two situations, mainly differing in development data. The results show that, for the approach that uses the Italic language family corpus to be the fine-tuning data and uses the development data where the L1 background is the same as the one of the test data, the GEC models improve clearly; however, the way that influences the models is not uniform, and varies by error types.

1 Introduction

Grammatical Error Correction (GEC) is a well-established Natural Language Processing (NLP) task, aiming to detect grammatical and other errors in texts and provide corrections automatically. With the increase in publicly available annotated learner corpora of English as Second Language (ESL), most GEC researchers have recognized the difference between ESL corpora and native-speaker corpora, and used ESL corpora to develop their systems to face ESL texts, which has led to quite impressive improvements. However, most of the works used the ESL corpora as a whole, ignoring the complexity and the difference inside the corpora in terms of various first language (L1) backgrounds. Among few GEC works with L1 information, a common method is to train / fine-tune a set of GEC models, and each model is trained by a corpus from one (and only one) specific L1 background (Mizumoto et al., 2011; Nadejde and Tetreault, 2019). The results of those works have shown its effectiveness, but it still has a shortcoming: the limited quantity of the training data. This kind of shortage refers to two aspects. Firstly, for most common freely available annotated datasets in English, only very limited ones supply the writers' L1 information. Secondly, it is far from enough for the utilization of the rare ESL corpus with L1 information: among all texts in the corpus, the texts from one specific L1 background can only be just a small fraction of the whole corpus, and other texts are not used when training the model for dealing with that specific L1. If the GEC system only uses a piece of data from the ESL corpus to be the training data, it means the remaining data, which is also proportionally the largest part of the corpus, does not contribute to the model's training at all. It is, by any means, a tremendous waste for the precious ESL corpus with L1 information.

*Corresponding author.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 922-933, Jinan, China, August 11-14, 2025.

To solve the problem of the data limitation, instead of using just one specific L1 corpus, this research aims to enlarge the amount of the dataset by using a set of corpora where writers' L1s are genetically related, or to say, are part of the same language family. A language family is a group of languages that share a common ancestral language or parental language (Coopmans, 1983). Languages in the same language family share similar language structures, related vocabularies, and even close culture background (Coopmans, 1983), making ESL learners in such L1s experience similar language transfer influence from their L1s to the target language, leading them to make similar biased errors, and consequently, leading to similar error patterns waiting to be learned by GEC models. Though the error patterns inside a language family can hardly be exactly the same as those in its specific member L1s, it is still a practical Second Language Acquisition (SLA) idea that can help to substantially enlarge the usable training data. Some research in other related areas has shown the practicability of using the similarity of the error patterns made by ESL learners from the same language family background. Nagata and Whittaker (2013), for example, has shown that the Indo-European language phylum can be correctly distinguished into Italic, Germanic, and Slavic language families by analyzing the error patterns of an ESL corpus where writers' L1s are their members, claiming that the relationship between languages that belong to the same language family is much stronger than those which are not. Thus, by reversing that idea, it is feasible to utilize the similarity between the language family and its inner members to enlarge the usable dataset.

In this work, we aim to find appropriate approaches to utilizing the similarity between the language family and its members to improve the GEC system. The structure of this paper is as follow: In Section 2, we introduce some related works for GEC based on L1 information. In Section 3, we illustrate two fine-tuning approaches for two situations, mainly different from the development data. In Section 4, we explain our experimental setup and designs of how to use a language family corpus. In Section 5, we discuss the results of experiments and make the analysis. Finally, we summarise this work and provide our future works in Section 6.

2 Related Works

2.1 GEC Based on L1 Information

Though the writer's L1 is an important additional information for the GEC task, due to the extremely limited ESL corpora with L1 information, the work based on this is quite limited. Most of the popular public ESL corpora have no L1 information of the ESL writers. The few exceptions include the First Certificate in English (FCE) corpus (Yannakoudakis et al., 2011) and the Lang-8 Learner Corpora (Mizumoto et al., 2011); some others, such as the Cambridge Learner Corpus (CLC) (Nicholls, 2003) and the International Corpus of Learner English (ICLE) (Granger, 2003) are not publicly available.

Mizumoto et al. (2011) developed a GEC system to correct Japanese (not English) texts. The system was based on the Statistical Machine Translation (SMT) paradigm, which considers grammatical error correction as a kind of translation from erroneous text to correct text. They used texts from two L1 (English and Mandarin) backgrounds and trained two GEC models, each of which was trained by a corpus from one (and only one) specific L1 background. This research proved that the GEC system performed better when the training and test data were from the same L1 background. This demonstrates that information about the writer's L1 is useful for GEC processing, while it still struggled with the lack-of-data issue.

To solve the data limitation problem, some of the research used specific model architecture (Rozovskaya et al., 2017; Chollampatt et al., 2016), making the results hard to reproduce to other models. Rozovskaya and Roth (2010a) proposed a common method which used the artificial errors to mimic the distribution of ESL errors, and used a native corpus (considered error-free) with artificial article errors to train a classifier to correct article errors for three L1s (Chinese, Czech, and Russian). The core idea for the artificial errors method is to use artificial erroneous sentences to train discriminative classifiers whose errors are generated at a rate that reflects the errors made by ESL learners to simulate error patterns in the real ESL corpus. They used an Averaged Perceptron classifier to build the L1-specified GEC systems to test the performance of correcting article errors for texts from their three chosen L1s, none of which has an article system. The experiments showed that, for all L1 groups, classifiers trained by corpus using the artificial errors method to introduce ESL errors outperformed the classifiers trained by the native English

Proceedings of the 24th China National Conference on Computational Linguistics, pages 922-933, Jinan, China, August 11-14, 2025.

corpus. Rozovskaya and Roth (2010b) expanded this work to five L1s (Chinese, Czech, Italian, Russian, and Spanish) and preposition errors, and showed similar results.

Nadejde and Tetreault (2019) used a Neural Machine Translation (NMT) model architecture and the transfer learning technique to solve the lack-of-data problem. They utilized L1 information and, furthermore, the proficiency level to improve the GEC model, using fine-tuning method to modify the performance of the pre-trained model by corpus with specific L1 and / or proficiency levels. They used a Long Short Term Memory (LSTM) model as their encoder and decoder model along with attention modules. The pre-trained model was an unnamed general-purpose NMT GEC system trained on 2M sentences written by both native speakers and ESL learners covering different topics and styles. The fine-tuning data were part of CLC, a large-scale non-public corpus. Specifically, the extracted sentences were grouped by the writers' L1, proficiency level, or L1-proficiency combination, and each group had at least 11,000 sentences, 8,000 sentences as the fine-tuning data, 1,000 as the development data, and 2,000 as the test data. The results showed that the fine-tuned models based individually on proficiency level or L1 were better than the baseline model fine-tuned by random CLC data on average by 2.1 and 2.3 $F_{0.5}$ score respectively, and the fine-tuned models based on both proficiency level and L1 outperformed all others, beating the baseline by 3.6 $F_{0.5}$ score on average. The shortage of this work includes: it only used a small fraction of the whole CLC corpus, which wasted the majority of the fine-tuning data; the pre-trained model and the fine-tuning data were all non-public, making the reproducibility extremely difficult.

3 Approaches to Utilizing Language Family Corpus

As mentioned, a language family is a group of languages that share a common ancestral language or parental language. Based on their relationship with each other, members of a language family can be grouped in the tree structure. For example, the Italic language family can be shown as Figure 1 (Atkinson and Gray, 2006). By utilizing the similarity between the language family as a whole and its inner members, the fine-tuning data for models that correct texts from specific L1 backgrounds can be enlarged dramatically.

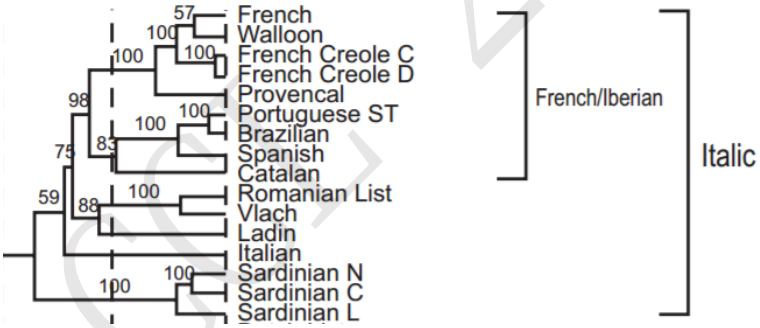

Figure 1: The Italic language family tree, from Atkinson and Gray (2006). While there is broad agreement about the structure, there are some minor differences among viewpoints. Atkinson and Gray (2006) inferred the above tree using Bayesian Markov Chain Monte Carlo in testing hypotheses about the age of the Indo-European language family. Values above each branch indicate uncertainty (posterior probability) in the tree as a percentage.

Compared with using a specific L1 corpus to fine-tune the GEC model, using a language family corpus is more complicated. Based on the composition of the development data, there are two hypothetical situations when using language family corpus, each with a corresponding approach.

The first situation can be described as: GEC for a text from the same L1 background. In this situation, the fine-tuning data (the language family corpus) is considered as a rough substitute dataset for the specific L1 corpus, which means the L1 background of the fine-tuning data is treated as the same one of the test data. Compared with the L1 corpus, the language family corpus increases its quantity while decreases its quality in terms of the similarity of the characteristics of errors and error patterns. However, based on the similarity of the language family and its member languages, this decrease is limited and acceptable. For

each piece of test data, the writer's L1 is clear, which means the L1 background of the development data can be exactly the same as the one of the test data. Based on that, there will be multiple GEC models, and the number of fine-tuning models equals the number of L1s involved in the test texts. In conclusion, in this situation, the fine-tuning data, the development data, and the test data are highly analogous (as all can be considered to be from the same *specific* L1 background), making the learning process of the NMT model better than the one with heterogeneous data.

An example to illustrate such a situation and the corresponding approach to utilizing language family corpus is as follows: suppose a text from French language background is the test data and the Italic language family corpus contains five L1 corpus including French L1 corpus. The GEC model chosen to correct the test data in this approach is the fine-tuning model using the Italic language family corpus as the fine-tuning data and the French L1 corpus as the development data. In this circumstance, the quality of the fine-tuning data is inferior to the French L1 corpus, since the GEC model may learn some error patterns which are less likely to be made by French ESL learners, but due to the genetic and linguistic link between French and other members in the Italic language family, this quality reduction is acceptable. On the other hand, the quantity of the fine-tuning data is a huge promotion: the fine-tuning data for using Italic language family corpus was approximately four times larger than using French L1 corpus only, for the other four related L1 corpora are joined together to form the fine-tuning data.

The second situation can be described as: GEC for a text where a writer's L1 is in a specific language family, but the exact L1 is unknown; or alternatively, GEC for texts where writers' L1s are in a specific language family, despite which the L1 exactly is. In this situation, the fine-tuning data (the language family corpus) is considered as one special L1 like "Italic L1". This new artificial L1 is considered to have the unified similarity of the error patterns as a whole, without distinguishing its internal differences. Based on that, when fine-tuning the GEC model in this situation, the development data should contain error patterns as the fine-tuning data involved, which means the L1 of the development data should also be the "Italic L1". In this approach to fine-tuning the GEC model, the development data and the test data are not exactly the same in terms of the L1 background and, thus, error patterns.

Actually, the similarity of language family and its members in this situation is projected to the similarity between the development data and the test data (the first situation is between the fine-tuning data and the test data): the fine-tuning data, the development data, and the test data are all considered from the "Italic L1" background, and the difference of the exact test data is ignored. What is more, there will be only one GEC model to be fine-tuned, because there is no need to change the development data, which greatly economizes the training cost and simplifies the system structure.

Here is an example to illustrate such a situation and its corresponding approach. Suppose the GEC system is correcting a text from French language background as the test data, and the Italic language family is as before. The GEC system is unclear about the L1 of the test text but knows that the writer's L1 is a member of the Italic language family, or alternatively, the GEC system just considers the L1 of the test text is the "Italic L1" and ignores the smaller differences in that "language". In this situation, the (one and only) GEC model is fine-tuned by the Italic language family corpus as the fine-tuning data and the development data. For such an approach, the quality of the fine-tuning data is considered the same as the test data, even if some aberration may be hidden and ignored in the test text, and the quantity of the fine-tuning data is roughly four times larger than using the French L1 corpus only. Further, the GEC system only needs to use one model to process all test data from the Italic language family corpus.

4 Experiment

4.1 Experimental Setup

4.1.1 GEC Systems and Model Training Details

We use the GEC architecture of Chollampatt and Ng (2018) as the basis for our L1-specific models. It is based on a fully convolutional encoder-decoder architecture with multiple layers of convolutions and an attention neural network module (Gehring et al., 2017). We choose Chollampatt and Ng (2018) as a well-established architecture that is often used as a baseline. While its performance has been exceeded by

Proceedings of the 24th China National Conference on Computational Linguistics, pages 922-933, Jinan, China, August 11-14, 2025.

more recent work, it still performs respectably, with 54.79% in $F_{0.5}$ score on the CoNLL-2014 test data and 57.47 GLEU score on the JFLEG test data (Napoles et al., 2017).[1]

For the pre-trained model, before training and correcting, the system splits rare words in parallel corpus and test texts into multiple frequent sub-words by using byte pair encoding (BPE) algorithm (Sennrich et al., 2016). After that, the system initializes the word embeddings for texts with pre-trained word embeddings by using the fastText tool (Bojanowski et al., 2017). When training, the model uses a negative log-likelihood loss function, and the parameters are optimized by Nesterov's Accelerated Gradient Descent (NAG) algorithm, using a simplified Nesterov's momentum formulation (Bengio et al., 2013). Fine-tuning models were initialized by the pre-trained model, and fine-tuned without any layer being frozen.

During the correction phase, given the erroneous source sentence, the system estimates the probability of target words obtained by a left-to-right beam search and retains the n best candidates at every decoding time step. At the end of the search, the correction hypothesis will be the candidate sequence of the highest sum of the log probabilities of all the candidate words in the beam. In addition, a word representation tool is used for the embedding layer of the model.

In terms of implementation, the model is built by Fairseq modeling toolkit (Ott et al., 2019), a publicly available PyTorch-based sequence to sequence modeling toolkit. The dimensions of embeddings are 500 for both source and target words. Both the source and target vocabularies contain about 30K most frequent BPE-based tokens, so does the word embeddings layer. The encoder and decoder both contain seven convolutional layers, with a convolution window of width 3. The dimension for each encoder and decoder layer is 1024. Dropout probability for the embeddings, convolution layers, and decoder output is 0.2 in all cases. The learning rate is 0.25, learning rate annealing factor is 0.1, and momentum value is 0.99; we use the same learning rate for initial training and fine-tuning. During decoding, a beam width of 12 is used. The evaluation to choose the best model when training is based on the $F_{0.5}$ score on the development data.

4.1.2 Dataset

We use the data extracted from the Lang-8 Learner Corpora and the NUCLE corpus to be the training data for the pre-trained model. Unlike Chollampatt and Ng (2018) who used a relatively rough method to extract source-target sentence pairs, we design a more precise method for pre-processing the Lang-8 Learner Corpora. The extraction process includes identifying English sentences, deleting system tags, replacing or deleting special characters, and deleting comments in target sentences. Extracted sentence pairs are about 2.5M, all to be the training data. The NUCLE corpus is chosen to be the training data (about 51.7K sentence pairs) and the development data (about 5.4K). Sentence pairs in the training data are discarded if the source sentence is the same as the corrected sentence.

The dataset for fine-tuning is from the FCE corpus. The FCE corpus contains writers from 16 different L1s. Among them, Catalan, French, Italian, Spanish, and Portuguese are in the Italic language family. The Italic language family corpus is the largest language family corpus in the FCE corpus, containing 13545 source-target sentence pairs within 554 texts. Therefore, we choose it to be the fine-tuning data of the experiments.

4.1.3 Evaluation

The evaluation tools are MaxMatch (M^2) Scorer (Dahlmeier and Ng, 2012)[2] and ERRor ANnotation Toolkit (ERRANT) (Bryant et al., 2017). M^2 Scorer was also used in the CoNLL-2014 Shared Task (Ng et al., 2014), giving precision, recall, and $F_{0.5}$ as its evaluation metrics. ERRANT is a rule-based automatic error-type provider, which aims to generate the error type of each difference between two inputs.

4.2 Experiment Design

As discussed before, there are two situations to utilize a language family corpus: GEC for texts from the same L1 background, and GEC for texts where the writers' L1s are in one language family, despite which

[1] As example comparison systems, on the CoNLL-2014 dataset, Zhao et al. (2019)'s copy architecture achieves a 61.15% $F_{0.5}$ score, Liu et al. (2021)'s approach with multiple hypotheses 63.7%, and the state-of-the-art system at writing, Rothe et al. (2021)'s fine-tuning of a multilingual T5 model, 68.87%. As of writing, Chollampatt and Ng (2018) still ranks in the top 15 in the leaderboard at https://paperswithcode.com/sota/grammatical-error-correction-on-conll-2014.

[2] http://www.comp.nus.edu.sg/~nlp/software.html

the L1 exactly is. Correspondingly, there are two approaches to implementing those situations, which we will refer to as *Approach 1*, and *Approach 2* respectively. We propose two experiments for a language family corpus to test how GEC systems perform when using Approach 1 and Approach 2.

For the Italic language family corpus used in the experiments, each L1 corpus inside the corpus is separated into five parts. The fine-tuning data for both two approaches is the combination of three parts of source-target sentence pairs from each L1 corpus in the Italic language family corpus. For the test data, one part of the sentence pairs is chosen from the remaining 2/5 part of each L1 corpus in the Italic language family corpus, forming the five L1 test data sets. For the development data, Approach 1 use data where the L1 of the writers is exactly the same as the one of the test data. More specifically, the remaining one part of the sentence pairs from each L1 corpus in the Italic language family corpus is fully used to be the development data of the corresponding L1 model. The development data used in Approach 2 is from the "Italic L1" background: it is randomly chosen 1/5 of the sentence pairs from each remaining one part of the L1 corpus and combine them as a whole to keep the quantity similar to Approach 1. Mentioned that Approach 1 generates five fine-tuning models, and each L1 model processes its related L1 test data, while Approach 2 generates only one fine-tuning model to process all five L1 test data in the Italic language family corpus.

As the baseline comparison, baseline models for both two approaches use fine-tuning data randomly extracted from the FCE corpus exclusive of the Italic language family corpus, and keep the quantity the same as the fine-tuning data of their corresponding experimental models. The development data and the test data for Approach 1 baseline models are the same as Approach 1 models. Approach 2 baseline model uses the test data the same as the Approach 2 model, and the method for extracting development data is the same as the Approach 2 model except the data used for extracting is changed to the same as the fine-tuning data of Approach 2 baseline model.

For clarifying, table 1 demonstrates the component of the training, development, and test data in different models. All L1 models using the same fine-tuning approach and language family corpus will be called in a group. For example, all L1 models using Approach 1 and Italic language family corpus will be called Approach 1 Italic language family group. Each cell shows the L1 background and the quantity of the data. For example, "Catalan $\times$ 3" means there are three parts of data where the L1 background of the data is Catalan, "Random in Italic" means there is one part of data which is randomly chosen from the Italic language family corpus, and "Random" means there is one part of data which is randomly chosen from the FCE corpus except the Italic language family corpus. Mentioned that the quantity of a part of "Random in Italic" and "Random" is the same as the average quantity of a part of other five L1 corpora as we described before.

	Approach 1 Italic language family group		**Approach 1 random corpus group**		**Test**
	Training	**Development**	**Training**	**Development**	
Catalan	Catalan $\times$ 3	Catalan	Random $\times$ 3	Catalan	Catalan
French	French $\times$ 3	French		French	French
Italian	Italian $\times$ 3	Italian		Italian	Italian
Spanish	Spanish $\times$ 3	Spanish		Spanish	Spanish
Portuguese	Portuguese $\times$ 3	Portuguese		Portuguese	Portuguese
	Approach 2 Italic language family group		**Approach 2 random corpus group**		**Test**
	Training	**Development**	**Training**	**Development**	
Catalan	Catalan $\times$ 3	Random in Italic	Random $\times$ 3	Random	Catalan
French	French $\times$ 3				French
Italian	Italian $\times$ 3				Italian
Spanish	Spanish $\times$ 3				Spanish
Portuguese	Portuguese $\times$ 3				Portuguese

Table 1: The L1 background and the quantity of training, development, and test data in different models.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 922-933, Jinan, China, August 11-14, 2025.

5 Results and Analysis

5.1 Approach 1 Groups

Overall Performance The performance for GEC systems with fine-tuning models by the Italic language family (Italic language family group) and relevant baseline systems (random corpus group) for Approach 1 is shown in Table 2.

	Approach 1 Italic language family group			Approach 1 random corpus group		
	Precision	**Recall**	**$F_{0.5}$**	**Precision**	**Recall**	**$F_{0.5}$**
Catalan	0.5433	0.3147	0.4744	0.5300	0.3069	0.4627
French	0.5470	0.3336	0.4850	0.5131	0.3107	0.4540
Italian	0.5229	0.2831	0.4471	0.4433	0.2583	0.3877
Portuguese	0.5081	0.3083	0.4498	0.4724	0.3043	0.4254
Spanish	0.5569	0.3041	0.4775	0.5210	0.2995	0.4539

Table 2: Results for the Approach 1 Italic language family group and the Approach 1 random corpus group.

As can be seen in Table 2, it is obvious that the performance of the Approach 1 Italic language family group was superior to its baseline group (Approach 1 random corpus group) in all three metrics: for precision, from 0.0133 better (Catalan) to 0.0796 better (Italian); for recall, from 0.0040 better (Portuguese) to 0.0248 better (Italian); for $F_{0.5}$ score, from 0.0117 better (Catalan) to 0.0594 better (Italian). Focusing on particular L1s, the improvement for the Italian case is the most obvious, which got the highest increase for all three metrics. What is more, the extent of improvement for the Spanish case and the Portuguese case was similar, which may be due to the closer linguistic relationship between them than other Italic language family members.

Detailed analysis We performed error type-specific performance for the Italic language family group and its baseline group for Approach 1 using ERRANT. From the ERRANT output, for each L1, we found the number of corrections for error types where the numbers of error cases are relatively large (all error cases in the Italic language family group and its baseline group are of size greater than 100), calculated the differences between two groups, and normalized them by the total errors, expressed as a percentage. Table 3 shows the calculation of the level of improvement.

	Catalan	**French**	**Italian**	**Portuguese**	**Spanish**	**Average**
SPELL	-6.25	-2.06	5.26	1.47	-0.50	-0.54
PREP	4.69	4.11	5.26	-1.47	0.00	2.17
DET	-6.25	-2.74	0.00	1.47	-0.50	-1.44
OTHER	-3.13	0.00	2.63	-1.47	-2.00	-0.90
VERB:FORM	1.56	0.00	2.63	-1.47	-0.50	0.18
PUNCT	-1.56	-1.37	-1.32	-1.47	-5.00	-2.71
ORTH	1.56	4.11	2.63	2.94	-1.00	1.63
VERB:TENSE	1.56	0.69	-2.63	0.00	0.50	0.18
NOUN:NUM	-1.56	0.00	1.32	1.47	5.50	2.17

Table 3: The level of improvement for results between the Italic language family group and its baseline group for Approach 1 for main error types.

As can be seen in Table 3, in terms of error type, the results performed variably. SPELL, PREP, and DET are the top three error types in quantity, and stood for three different trends.

DET stands for the first trend, which is generally negative. Only Portuguese had a positive change (1.47), and others got no change or a negative one (from -6.25 to 0), making the general change negative

Proceedings of the 24th China National Conference on Computational Linguistics, pages 922-933, Jinan, China, August 11-14, 2025.

(-1.44 on average, using total difference normalized by the total number of texts), which means using language family corpus may harm the performance of the GEC fine-tuning models. A similar case is PUNCT (all negative change, -2.71 on average). Take the following sentence as an example:

(1) The exams we had to do next day were there .

The sentence is written by a Catalan ESL learner, and the Catalan model got the lowest level of improvement when using Approach 1. Before *next* there should be a *the* which is missing in the sentence. The Catalan model in Approach 1 Italic language family group failed to detect this error, as with the one in L1-SPECIFIC, while its baseline model (the one in Approach 1 random corpus group, same as below) successfully detected this error and added the missing article correctly.

The second trend is generally positive, as PREP showed. For PREP, all but Portuguese (-1.47) got non-negative change (from 0 to 5.26), making the general change positive (2.16 on average), which means using language family corpus may help the performance of the GEC fine-tuning models. Similar situations include ORTH (one negative and four positive, 1.63 on average) and NOUN:NUM (one negative and three positive, 2.17 on average). Take the following sentence as an example:

(2) So , when Peter , who has been a friend of mine since we were children invited me for a pizza , I accepted without thinking at the consequences .

The sentence is written by an Italian ESL learner, and the Italian model got the highest level of improvement when using Approach 1. After *think* the preposition *in* is misused and should be replaced by *of* in the sentence. The Italian model in Approach 1 Italic language family group successfully detected this error and replaced the preposition correctly, while its baseline model and the model in L1-SPECIFIC failed to detect this error.

SPELL stands for the third trend, which is generally neutral. For SPELL, two L1s (Italian and Portuguese) had a positive change (5.26 and 1.47), and others had a negative change (from -6.25 to -0.50), making the general change neutral (-0.54 on average), which means using language family corpus make no harm nor help for the performance of the GEC fine-tuning models. Similar situations can be found in VERB:FORM (two negative and two positive, 0.18 on average) and VERB:TENSE (one negative and three positive, 0.18 on average). Take the following sentences as examples:

(3) a. I like very much Danny Brook , and his unaspected absence caused on me a big sadness .
b. Chairs are not only part of the forniture , they are also objects of decoration and you can even find some in museums as art symbols .

The sentences 3a and 3b are written by an Italian and a Catalan ESL learner respectively, and the Italian and the Catalan model got the highest and the lowest level of improvement when using Approach 1 respectively. In the sentence 3a, there is a misspelling word *unaspected* which should be replaced by *unexpected*. The Italian model in Approach 1 Italic language family group successfully detected this error and replaced the misspelling word correctly, while its baseline model and the model in L1-SPECIFIC, though they detected the error, failed to replace the misspelled word to the corrected one but proposed *unspectacular* and *unattencted* respectively. In the sentence 3b, there is a misspelled word *forniture* which should be replaced by *furniture*. The Catalan model in Approach 1 Italic language family group and the one in L1-SPECIFIC failed to detect that error, while the baseline model detected this error and replaced the misspelled word correctly.

In terms of L1, Catalan got two major (the magnitude of level of improvement $\geq$ 4.00) negative influences by DET (-6.25) and SPELL (-6.25) while one major positive influence by PREP (4.69), making the Catalan GEC model got the least improvement. Italian got two major positive influences by PREP (5.26) and SPELL (5.26), and no major negative influence, making the Italian GEC model got the greatest improvement. Compared with Spanish and Portuguese, though the overall performance is similar, the inside performance for different error types is complex. On one hand, generally speaking, the overall trend for different error types are similar: only three error types of all error types have reverse positive / negative change, which is the least one compared with any two L1s in the Italic language family. On the

Proceedings of the 24th China National Conference on Computational Linguistics, pages 922-933, Jinan, China, August 11-14, 2025.

other hand, the level of improvement for the two L1s is different. Spanish got one major negative change by PUNCT (-5.00) and one major positive change by NOUN:NUM (5.50), while Portuguese got no major negative or positive change. All the above may be caused by the similar error patterns inside the Spanish and Portuguese L1 corpora while different in quantity, making the results of the two GEC models have a similar trend for error types while different in the level of improvement.

5.2 Approach 2 Groups

The performance for the GEC system with the fine-tuning model by the Italic language family (Italic language family group) and the relevant baseline system (random corpus group) for Approach 2 is shown in Table 4.

	Approach 2 Italic language family group			**Approach 2 random corpus group**		
	Precision	**Recall**	**$F_{0.5}$**	**Precision**	**Recall**	**$F_{0.5}$**
Catalan	0.5571	0.3108	0.4809	0.5282	0.2896	0.4534
French	0.5517	0.3107	0.4776	0.5310	0.3248	0.4712
Italian	0.4872	0.2748	0.4220	0.4755	0.2810	0.4177
Portuguese	0.4860	0.2747	0.4212	0.5018	0.2708	0.4287
Spanish	0.5334	0.3080	0.4653	0.5400	0.3048	0.4678

Table 4: Results for the Approach 2 Italic language family group and the Approach 2 random corpus group.

Table 4 illustrates that, there is no clear improvement in general for GEC systems with the Italic language family fine-tuned mode for Approach 2. Compared with its baseline group, Approach 2 Italic language family group was from 0.0158 worse (Portuguese) to 0.0289 better (Catalan) for precision, from 0.0141 worse (French) to 0.0212 better (Catalan) for recall, and from 0.0075 worse (Portuguese) to 0.0275 better (Catalan) for $F_{0.5}$ score. The most obvious improvement of Approach 2 Italic language family group is the case of Catalan (0.0275 better for $F_{0.5}$). The other cases were variable, but the magnitude of change (either positive or negative) in $F_{0.5}$ was all less than 0.01, which means there was roughly no effect for the GEC system in using the Italic language family corpus for fine-tuning using Approach 2. All the above demonstrates that Approach 2 is not an effective method to use the similarity between language family and its inside members, or to say, it is not a good approximation of Approach 1.

Compared with the Italic language corpus group for Approach 1 and Approach 2, generally speaking, the performance of Approach 1 is superior to Approach 2. Except for Catalan (0.0065 worse), all cases in the Italic language corpus group for Approach 1 are better than the one for Approach 2 (from 0.0074 better to 0.0286 better) for $F_{0.5}$ score. Such superior demonstrates that the homogeneity between the development data and the test data is much more important than the one between the fine-tuning data and the test data. The main difference between Approach 1 and Approach 2 is the component of the development data. The development data for Approach 1 is the corpus where writers' L1s are exactly the same as those of the test data, making the GEC system use five models to deal with texts from different L1 backgrounds. While the only "L1" for the development data for Approach 2 was "Italic L1", which is an approximation of the exact test texts' L1s. The inferior performance of the Italic language family group for Approach 2 compared with the one for Approach 1 has shown that, the development data should be as similar as possible to the test data, or it can harm the performance of the model.

Same as Approach 1, the level of improvement for main error types is shown in Table 5. As can be seen, the average performance for each error type is with little difference. Except for DET (-2.17), no error type improves or influences more than one. The worst case is DET (-2.17), and the best one is PUNCT (0.90). The ordinary level of improvement illustrates that the fine-tuning using Approach 2 achieved little improvement.

	Catalan	French	Italian	Portuguese	Spanish	Average
SPELL	1.56	1.37	2.63	-5.88	-1.00	-0.18
PREP	7.81	-1.37	0.00	-2.94	1.50	0.72
DET	-10.94	-3.42	0.00	2.94	-1.00	-2.17
PUNCT	7.81	-8.90	1.32	4.41	4.50	0.90
VERB:FORM	0.00	1.37	1.32	0.00	0.00	0.54
OTHER	-1.56	-2.05	1.32	-4.41	2.00	-0.36
ORTH	1.56	2.74	0.00	1.47	-2.00	0.36
NOUN:NUM	1.56	2.05	-1.32	-7.35	2.00	0.36
VERB:TENSE	-1.56	0.00	-1.32	2.94	1.50	0.54

Table 5: The level of improvement for GEC systems between the Italic language family group and its baseline group for Approach 2 for main error types.

6 Conclusion

In this research, we utilized the similarity between the language family as a whole and its inner member languages to propose methods of using language family corpus to enlarge the fine-tuning data. We proposed two different approaches of using language family corpus for two different situations and tested their performance by using the Italic language family corpus. Experiments showed that, for the approach that uses Italic language family corpus to be the fine-tuning data and uses development data where the L1 background is the same as the one of the test data, the performance was substantially better than using equivalent amounts of fine-tuning data randomly chosen. However, the way that influences the models is not uniform, and varies by error types: DET errors are negatively influenced, PREP errors positive, and errors SPELL neutral. Besides, it was the more effective approach to utilizing a language family corpus compared with the approach that uses the development data from the "Italic L1" background.

However, there still remain two issues waiting to be solved. First of all, though compared with the method of fine-tuning GEC model by specific L1 corpus, the level of utilization of the ESL corpus with L1 information is increased when using the language family corpus to enlarge the fine-tuning data, a great number of texts which do not belong to the language family corpus are still not utilized, which still makes a waste of training data. Secondly, this method can only process test texts where writers' L1s are in the larger language family that has available data; when facing texts from other L1 backgrounds, this method cannot be used, which greatly restricts its application. Thus, in the future works, we aim to investigate different approaches to utilize L1 information which can solve those problems.

References

Quentin D Atkinson and Russell D Gray. *How old is the Indo-European language family? Illumination or more moths to the flame*, volume 91, page 109. Cambridge: McDonald Institute for Archaeological Research, 2006.

Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628. IEEE, 2013.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 06 2017. ISSN 2307-387X. doi: 10.1162/tacl_a_00051. URL https://doi.org/10.1162/tacl_a_00051.

Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1074. URL https://aclanthology.org/P17-1074.

Shamil Chollampatt and Hwee Tou Ng. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, February 2018.

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*, pages 1901–1911, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1195. URL https://www.aclweb.org/anthology/D16-1195.

Peter Coopmans. Language universals and linguistic typology. *Journal of Linguistics*, 19(2):455–473, 1983.

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*, pages 568–572, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/N12-1067.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org, 2017.

Sylviane Granger. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, 37(3):538–546, 2003. ISSN 00398322. URL http://www.jstor.org/stable/3588404.

Zhenghao Liu, Xiaoyuan Yi, Maosong Sun, Liner Yang, and Tat-Seng Chua. Neural quality estimation with multiple hypotheses for Grammatical Error Correction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5452, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.429. URL https://aclanthology.org/2021.naacl-main.429.

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*, pages 147–155, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I11-1017.

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)*, pages 27–33, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5504. URL https://aclanthology.org/D19-5504.

Ryo Nagata and Edward Whittaker. Reconstructing an Indo-European family tree from non-native English texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1147, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1112.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2037.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1701.

Diane Nicholls. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581, 2003.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. A simple recipe for multilingual Grammatical Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.89. URL https://aclanthology.org/2021.acl-short.89.

Alla Rozovskaya and Dan Roth. Training Paradigms for Correcting Errors in Grammar and Usage. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 154–162, Los Angeles, California, June 2010a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N10-1018.

Alla Rozovskaya and Dan Roth. Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 961–970, Cambridge, MA, October 2010b. Association for Computational Linguistics. URL https://aclanthology.org/D10-1094.

Alla Rozovskaya, Dan Roth, and Mark Sammons. Adapting to learner errors with minimal supervision. *Computational Linguistics*, 43(4):723–760, December 2017. doi: 10.1162/COLI_a_00299. URL https://aclanthology.org/J17-4002.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

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*, pages 180–189, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1019.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving Grammatical Error Correction via pre-training a copy-augmented architecture with unlabeled data. 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 156–165, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1014. URL https://aclanthology.org/N19-1014.