Syntactic-knowledge-based Grammar Error Correction for Second Language English Speakers

Taaha Kazi **Sharvi Tomar** **Shresta Bangaru** **Aditi Wikhe**
tnkazi2@illinois.edu, stomar2@illinois.edu,
bangaru2@illinois.edu , aditiw2@illinois.edu

Abstract

In this work, we have built Grammar Error Correction (GEC) models for second language speakers of English and have proposed two novel methodologies to improve their performance. Firstly, we propose a method that leverages syntactic information, such as part-of-speech (POS), to correct erroneous English sentences. Secondly, we propose a methodology to augment any training corpus based on the grammatical errors second language speakers make. This augmented training data is then used to train the models. Both these methods improve precision and recall scores for Grammar Error Correction.

1 Introduction

Improving GEC technology is essential for communication in various forms such as video captioning, conference transcripts, and online discussion forums. Potential consumers of such a technology would include educational institutions, global companies, and social media content creators who may want to reach large audiences. Better GEC has the potential to bring together people of different linguistic and cultural backgrounds in collaborative and professional environments.

Our approach aims to build on current GEC algorithms by incorporating Part-Of-Speech (POS) tagging and speakers' first language mapping. In previous studies, these have been shown to be closely tied to the type and frequency of grammatical errors made by the individual when writing in English.

2 Background and Related Work

2.1 Effect of First Language on English Grammar

From a linguistics and psychology perspective, the current understanding is that a person's first language significantly affects their thought process and sentence construction in English (Hua).

In a 2020 study about Grammar Error Correction for second language learning, researchers demonstrate the feasibility of using an individual's first language as a predictor for spelling, punctuation, and grammatical errors made in a second language; the first language itself can indicate errors with 15 to 25 per cent accuracy, which is significant for such little amount of information.

A similar study at MIT affirms this and additionally links the grammar mistakes made per first language group with the POS tagging results of the sentence (Berzak et al., 2016). This makes the first language label and POS tags of a sentence both compatible indicators of potential grammatical errors.

From a GEC implementation perspective, (Nadejde and Tetreault, 2019) built a neural GEC system for Second Language speakers. The adapted GEC models were built using fine-tuning and were compared on the CoNLL 2014 test set. Five proficiency levels and twelve different languages were explored in their research and they had three different scenarios: adapting only to the proficiency level, only to the first language and simultaneously addressing both aspects. The methodology showed improvement in all 3 scenarios.

3 Methodology

3.1 Preliminary Analysis

To get a better understanding of the data we were working with, we visualized the distribution of grammatical errors made by native speakers of Spanish, French, and Korean, refer to Appendix 8. The error categories are explained in section 3.2.1. Spanish and French are both Romance languages with a lexical similarity of 75 per cent, each with a 30 to 50 per cent similarity to English. Thus, the types and frequencies of mistakes that the native speakers of these languages make are similar.

Korean, on the other hand, is part of the Altaic language family and is very different from English. Native Korean speakers tend to leave out the determinant when writing in English, which makes sense because the Korean language does not use determinants such as “a”, “an”, and “the”. This preliminary analysis allowed us to establish the native language as a strong indicator of errors made in English writing.

3.2 Data Description

3.2.1 Error Categories

The error categories have been generated using ERRANT evaluation metrics (Bryant et al., 2017). The 25 main error categories are mentioned in Figure 1 and most of them can be prefixed with ‘M:’(Missing), ‘R:’ (Replacement) or ‘U:’(Unnecessary) edit to allow further fine-grained evaluation.

3.3 Data preprocessing

3.3.1 Sampling using error suppression

In the error file, each sentence is followed by any number of annotations which indicate grammatical error corrections. Some of the error labels are overcounted with regard to the original error distribution for that first language. We implemented error suppression to manually correct certain annotations in order to bring the file’s error count to a target distribution. To do this, we parsed the raw data file and stored the sentences and annotations by index in a dictionary data structure. Given a target distribution of errors mapped to frequency, our script used string operations to manually correct certain errors to achieve the desired distribution.

3.4 System Architecture

3.4.1 Parts-of-Speech Infused Embedding

The task of incorporating POS information is carried out in two steps: Firstly, we fine-tuned a large-language model(BERT base-cased) on a POS classification (or labelling) task as depicted in Figure **??** and extract the representations from BERT. We then fuse the extracted representations with each layer of the encoder and decoder of the GEC model through attention mechanisms and train the GEC on sentence pairs of the learner’s corpora. The model architecture we employed to infuse the syntactic information into our GEC model has been adopted from the work of (Zhu et al., 2020) and has further been explained in 3.4.2.

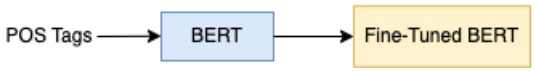

Figure 3: Procedure of Fine-tuning of BERT

The visualisation of contextual token embeddings extracted from BERT can be found in Appendix 8. Since embeddings carry not only syntactic but also lexical information, hence we do not see clear clusters of words or of POS tags but we presume that the explicit injection of syntax structure would help the embeddings to get transformed in a way that helps in the downstream tasks of GEC.

3.4.2 Transformer-based model

Our approach is inspired by (Zhu et al., 2020) and BERT-FUSE for GEC (Masahiro Kaneko and Inui, 2020); here we have used the BERT-fuse model for the GEC task. The motivation behind this approach is to incorporate the embeddings from a pre-trained language model into another pretrained model. The inner workings of the model as described in the paper are as follows: Given an input sentence X; The BERT model first encodes the input sentence which is represented as B(X). The GEC model then encodes both X and B(X) as inputs. For each layer, the Encoder model then applies a separate attention mechanism to both these embeddings and sums the result. Mathematically the process can be written as:

$$\tilde{h_i} = \frac{1}{2}\left(A_h\left(h_i^{l-1}, \mathbf{H}^{l-1}\right) + A_b\left(h_i^{l-1}, \mathbf{B}^{l-1}\right)\right) \tag{1}$$

where the hidden representation of each layer of the encoder is represented as $\tilde{h}_i$; A_b and A_h represent the attention models.

The output of each layer is then processed in a Feed Forward Network similar to the transformer model.(Vaswani et al., 2017)

The Decoder follows a similar approach which can be represented mathematically as:

$$\tilde{s}_i^l = \frac{1}{2}\left(A_h\left(\hat{s}_i^{l-1}, \mathbf{H}^{l-1}\right) + A_b\left(\hat{s}_i^{l-1}, \mathbf{B}^{l-1}\right)\right) \tag{2}$$

where the hidden representation of each layer of the decoder is represented as $\hat{s}_i^l$. Diagrammatically the system architecture is represented in Figure 2.

Code	Meaning	Description / Example
ADJ	Adjective	*big → wide*
ADJ:FORM	Adjective Form	Comparative or superlative adjective errors. *goodest → best, bigger → biggest, more easy → easier*
ADV	Adverb	*speedily → quickly*
CONJ	Conjunction	*and → but*
CONTR	Contraction	*n't → not*
DET	Determiner	*the → a*
MORPH	Morphology	Tokens have the same lemma but nothing else in common. *quick (adj) → quickly (adv)*
NOUN	Noun	*person → people*
NOUN:INFL	Noun Inflection	Count-mass noun errors. *informations → information*
NOUN:NUM	Noun Number	*cat → cats*
NOUN:POSS	Noun Possessive	*friends → friend's*
ORTH	Orthography	Case and/or whitespace errors. *Bestfriend → best friend*
OTHER	Other	Errors that do not fall into any other category (e.g. paraphrasing). *at his best → well, job → professional*
PART	Particle	*(look) in → (look) at*
PREP	Preposition	*of → at*
PRON	Pronoun	*ours → ourselves*
PUNCT	Punctuation	*! → .*
SPELL	Spelling	*genectic → genetic, color → colour*
UNK	Unknown	The annotator detected an error but was unable to correct it.
VERB	Verb	*ambulate → walk*
VERB:FORM	Verb Form	Infinitives (with or without "to"), gerunds (-ing) and participles. *to eat → eating, dancing → danced*
VERB:INFL	Verb Inflection	Misapplication of tense morphology. *getted → got, fliped → flipped*
VERB:SVA	Subject-Verb Agreement	*(He) have → (He) has*
VERB:TENSE	Verb Tense	Includes inflectional and periphrastic tense, modal verbs and passivization. *eats → ate, eats → has eaten, eats → can eat, eats → was eaten*
WO	Word Order	*only can → can only*

Figure 1: 25 Error Categories from ERRANT (Bryant et al., 2017)

4 Experimental Setup

4.1 Data

We used the FCE-train (Yannakoudakis et al., 2011) data for building our initial L2 language error distribution set. For training the GEC model, we use the WI-train (Yannakoudakis et al., 2011) dataset and the WI-dev as a development set. For training the individual GEC models for the L2 speakers, we apply the under-sampling mentioned in section 3.3.1 to the train set. For our experiments, we have applied the technique for Russian, Spanish and Korean L2 speakers. For testing and evaluation, we use the FCE-test (Yannakoudakis et al., 2011) data. For evaluating the L2-specific models, we use the test examples tagged with the respective L2 tag.

4.2 Models

For carrying out a comparative study of the impact of our proposed techniques we train the following models. First, we fine-tune the gigaword model from (Kiyono et al., 2019) on the training dataset as our baseline. This model is referred as the Vanilla Transformer model in further sections. Second, for evaluating the L2 specific models, we apply the under-sampling technique to the training data

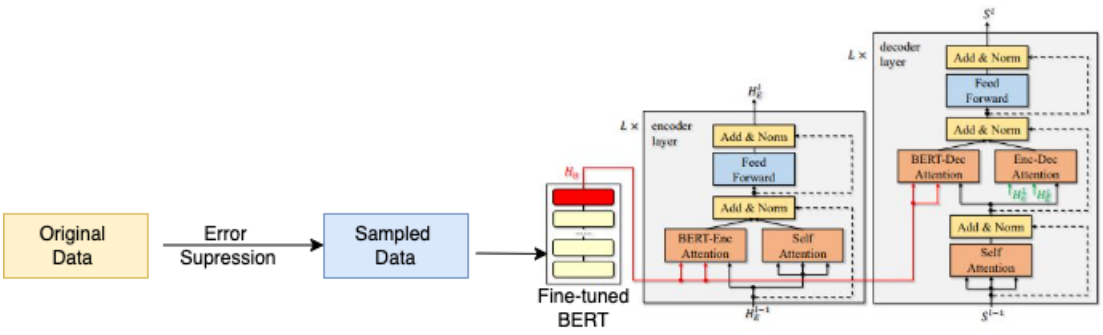

Figure 2: System Flow adopted from BERT-FUSE (Zhu et al., 2020)

Model	Train Data	Test Data	Precision	Recall	0.5 F-score
Vanilla Transformer	Whole	Russian FCE	0.5418	0.4969	0.5322
Vanilla Transformer	Russian Distribution-Sampled	Russian FCE	0.5512	**0.5123**	0.5429
POS-BERT-Fused	Whole	Russian FCE	0.5608	0.5092	0.5497
POS-BERT-Fused	Russian Distribution-Sampled	Russian FCE	**0.5942**	0.5031	0.5734
Vanilla Transformer	Whole	Korean FCE	0.3654	**0.4854**	0.3844
Vanilla Transformer	Korean Distribution-Sampled	Korean FCE	0.3654	**0.4854**	0.3844
POS-BERT-Fused	Whole	Korean FCE	0.3922	0.4781	0.4068
POS-BERT-Fused	Korean Distribution-Sampled	Korean FCE	**0.4177**	0.4818	0.4291
Vanilla Transformer	Whole	Spanish FCE	0.4707	0.4237	0.4605
Vanilla Transformer	Spanish Distribution-Sampled	Spanish FCE	0.4878	0.4403	0.4775
POS-BERT-Fused	Whole	Spanish FCE	**0.4918**	**0.4253**	0.4769
Vanilla Transformer	Whole	FCE	0.4798	**0.4359**	0.4703
POS-BERT-Fused	Whole	FCE	**0.4986**	0.4295	0.483

Table 1: Results

and fine-tune the model, similar to fine-tuning the baseline. Third, to evaluate the impact of Parts of Speech Embeddings, we train the models using the BERT-FUSE architecture, as mentioned in the 3.4.2. The BERT model used for our experiments was from (vblagoje).

4.3 Evaluation

For evaluation, we used GEC evaluation data on FCE-test. We used ERRANT evaluation metrics(Bryant et al., 2017).

5 Results

The detailed results achieved are mentioned in Table 1.

From Table1, we observe that sampling the data with our proposed strategy of error suppression leads to improvement in the evaluation metrics of GEC namely precision and recall. This general trend is observed in all three languages.

The incorporation of parts-of-speech embedding has given an average 5-point boost in the precision scores while maintaining recall scores. This result is seen across on all three languages we carried out our experiment. Also, in the last row of 1, i.e. models trained without sampling, the POS-BERT-Fused model improves precision while maintaining recall scores.

Analysis of the error types generated from errant shows that the models trained on the sampled data, predict higher number of Punctuation, Determinant and Preposition Corrections. Apart from these three error categories, the other error categories relatively remained the same.

Also, although the BERT-Fused models proposed fewer edits as compared to the vanilla transformer models, the edits they proposed were more frequently correct. This explains the jump in precision with the BERT-Fuse models.

6 Challenges and Limitations

There were a few challenges our team faced with the resources that were available. We had limited data for grammatical errors tagged by native language, as well as error overcounting in the annotations file. Part of the strategy to circumvent this included using error suppression to achieve a target error distribution per native language.

While using Google Colab, we encountered issues with disk space and runtime. Due to the high volume of data we were working with, we split up the various iterations of running the two models among the team members. Some of the scripts had to be run multiple times to obtain the final trained model, as the data was occasionally lost due to disk space limits on Colab. Limited GPU availability resulted in the runs being spaced out over the course of a week as well.

7 Conclusion and Future Work

In this paper, we proposed two methods to improve GEC performance for second-language English speakers, incorporating part-of-speech information into the GEC model and a novel sampling technique to augment any training set. Both these methods have shown to improve the performance of the GEC models measured by precision, recall and F1-score metrics. For future work, we aim to build

a re-ranking module that can work with a generic GEC model to re-rank the hypothesis generated by the model based on the first language of the speaker.

8 Contributions

All 4 authors contributed to writing this paper. Additional individual contribution of each author is described in Table 2.

Author	Contribution
Taaha Kazi	Model Training
	Evaluation Scripts
	Data Preprocessing
Sharvi Tomar	BERT fine-tuning,
	Related Work
	Model training (Russian)
Shreshta Bangaru	Data Prepocessing,
	Preliminary Analysis
	Model training (Spanish)
Aditi Wikhe	Inference Scripts
	Model training (Korean)

Table 2: Contributions

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Edward Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China. Association for Computational Linguistics.

Shun Kiyono Jun Suzuki Masahiro Kaneko, Masato Mita and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Maria Nadejde and Joel Tetreault. 2019. 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

vblagoje. bert-english-uncased-finetuned-pos.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. 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. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation.

A Appendix

A.1 Preliminary Analysis

Spanish and French are Romance languages. They have a lexical similarity of 75%. Each of them has about a 30-50% similarity to English. So the types and frequencies of mistakes that Spanish and French speakers make when writing in english is pretty similar.

Korean is part of the Altaic family of languages. It is not very similar to English. You can see that people whose first language is Korean tend to commonly make the error of missing a Determinant when they speak in English. A determinant is like “a”, “an”, “the” - article words Korean actually does not have determinants so it makes sense why this is the case.

This is just a small subset of the data we have but we wanted to visualize the English error distribution of various non-native speakers to show that (1) there are significant differences between the languages and (2) it will be feasible to train the model using this data.

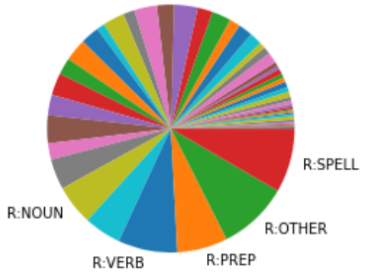

Figure 4: Error Distribution of Spanish Speakers

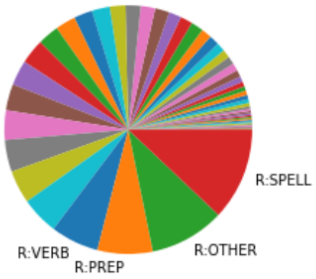

Figure 5: Error Distribution of French Speakers

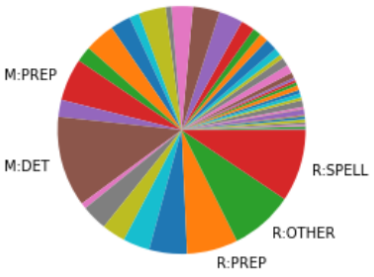

Figure 6: Error Distribution of Korean Speakers

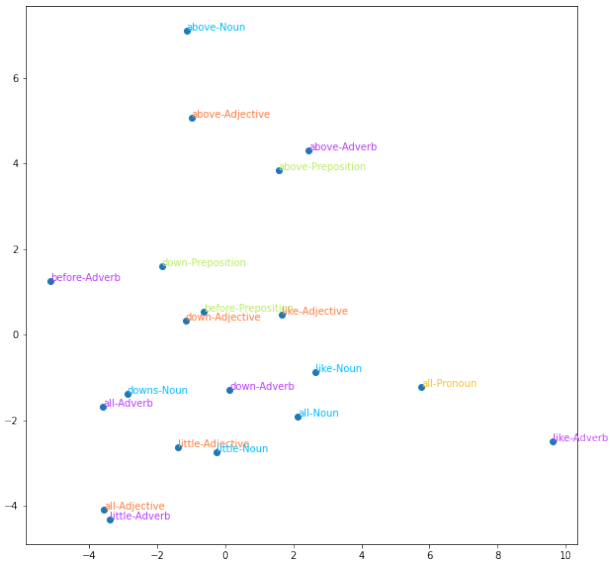

Figure 7: Embeddings from pretrained BERT

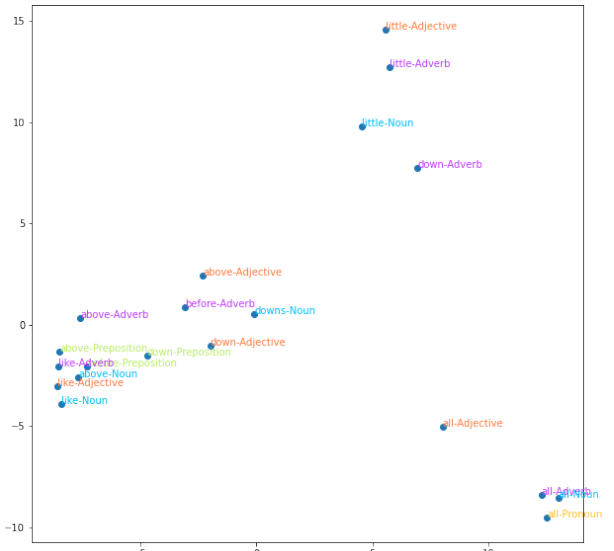

Figure 8: Embeddings from fine-tuned BERT

A.2 BERT Embedding Visualization

The idea of visualization of embeddings is to understand the difference in embeddings of the same words being used as different parts of speech in different sentences obtained from pre-trained BERT Figure 7 and the BERT which has been fine-tuned for the POS labelling task Figure 8. Each embedding is a high-dimensional vector of length 768 which has been plotted in 2-D using the t-SNE method.