# LLM-based Code-Switched Text Generation forError Correction

Tom Potter University of Manchester [thomas.potter@postgrad.manchester.ac.uk](mailto:
Zheng Yuan King's College London [zheng.yuan@kcl.ac.uk](mailto:zheng.yuan@kcl.ac.uk

## Abstract

With the rise of globalisation, code-switching (CSW) has become a ubiquitous part of multilingual conversation, posing new challenges for natural language processing (NLP), especially in Grammatical Error Correction (GEC). This work explores the complexities of applying GEC systems to CSW texts. Our objectives include evaluating the performance of state-of-the-art GEC systems on an authentic CSW dataset from English as a Second Language (ESL) learners, exploring synthetic data generation as a solution to data scarcity, and developing a model capable of correcting grammatical errors in monolingual and CSW texts. We generated synthetic CSW GEC data, resulting in one of the first substantial datasets for this task, and showed that a model trained on this data is capable of significant improvements over existing systems. This work targets ESL learners, aiming to provide educational technologies that aid in the development of their English grammatical correctness without constraining their natural multilingualism.

‘‘‘latex

# 1 Introduction

Code-switching (CSW), the practice of fluidly alternating between two or more languages in conversation, has become commonplace in recent years. This linguistic phenomenon, emerging as a natural consequence of multilingualism, is now widely accepted in social and professional settings (Yow et al., 2018). Many works have highlighted the utility and cultural importance of CSW in general conversation (Beatty-Martínez et al., 2020; Falbo and LaCroix, 2021). Further research indicates that these advantages extend to language learning, with CSW offering many pedagogical benefits. These include increasing students' access to content and improving their confidence. Nguyen et al. (2022) discuss the mechanisms for this, where students use a familiar language to grasp foreign, complex concepts. CSW can also serve as a scaffolding tool, helping to bridge gaps in a student's comprehension of a language and enabling them to build upon existing knowledge. These benefits reduce the barriers between a student and their target language and help promote a learning environment conducive with active exploration and deeper understanding. Therefore, it is essential that English as a Second Language (ESL) learners are not penalised for expressing their cultural identity through CSW.

Grammatical error correction (GEC) is the task of automatically detecting and correcting errors in text. Research on GEC for CSW text remained largely unexplored. Chan et al. (2024) were the first to demonstrate that exposing a sequence-tagging GEC model to CSW text during the training process improves performance compared to a monolingual system. However, further work is essential to ensure language technology is inclusive and reflective of real-world linguistic practices. Figure 1 shows two examples of CSW from our target population with their grammatical corrections.[1]

Despite significant advancements in GEC in recent years, a gap persists in addressing CSW texts, with monolingual GEC datasets labelling CSW as a type of error (Nguyen et al., 2022). There are several reasons for this, the most prominent being the scarcity of high-quality training data, a problem that plagues monolingual GEC systems. The unique linguistic features of CSW, including its variable syntax, semantics and pragmatics, add additional complexity to this task. Monolingual seq2seq GEC models, e.g. T5 (Rothe et al., 2021), struggle with CSW text as they fail to represent the non-English inputs, resulting in their inability to output the CSW text. On the other hand, multilingual seq2seq models and edit-based GEC models like GECToR (Omelianchuk et al., 2020) can handle CSW text but struggle with the ambiguity present at language switching points. This ambiguity challenges the models' ability to accurately correct the text.

IMAGE NOT PROVIDED

Figure 1: Examples of GEC in ESL learner language.

This paper aims to bridge this gap. Firstly, to address the data scarcity issue, we propose a method for generating high-quality synthetic CSW GEC data, using which we produce, to our knowledge, one of the first substantial datasets labelled for this task.[2] Secondly, we train a token classification-style GEC system, tailored to correct errors in texts produced by ESL learners. This demographic is significant for our study as they not only present consistent CSW patterns but also stand to benefit greatly from a GEC system capable of handling CSW text.

‘‘‘

‘‘‘latex

# 2 Data

## 2.1 Genuine CSW GEC Dataset

One of the only datasets labelled for GEC which does not remove CSW text is the Lang-8 dataset (Mizumoto et al., 2013), sourced from the Lang-8 language learning platform. This dataset, when filtered to contain entries where CSW is present, offers a foundation of authentic data, and comprises 5,875 pairs of ungrammatical and corrected sentences across 6

[1] The definition of CSW is a subject of ongoing debate. Throughout this work, we use the term CSW to refer specifically to the type of language mixing exhibited by ESL learners.

[2] This dataset is available on GitHub.

CSW language pairs: English-Japanese (81.9%), English-Korean (13.0%), English-Traditional Chinese (3.4%), English-Russian (1.2%), English-Thai (0.5%) and English-Arabic (0.1%).

The crowd-sourced nature of Lang-8 required manual validation to ensure accuracy. We tasked an annotator with the responsibility of verifying the original corrections in the dataset, as well as combing for missed errors, incorrect annotations and over-annotations.

## 2.2 Synthetic CSW GEC Data Generation

Given the small size of the available CSW GEC dataset, we introduced a 2-step approach to synthetic CSW GEC data generation. First, we generated grammatically correct CSW sentences. This is followed by the introduction of errors.

### 2.2.1 Step 1: CSW Text Generation

To generate diverse CSW texts without relying on existing corpora or inaccurate alignment algorithms, we leveraged the strong general knowledge of Large Language Models (LLMs).

We demonstrated that OpenAI's GPT-3.5 (Brown et al., 2020) can create high-quality CSW sentences when shown examples of authentic utterances.

Along with genuine CSW texts, we supplied a one-shot example of how to use the switching styles of an existing CSW text to generate a new sentence.[3]

Comparison of Synthetic CSW Text We used several CSW metrics to quantify the qualities of CSW texts: Code Mixing Index (CMI) (Gambäck and Das, 2016), Multilingual Index (M-Index) (Barnett et al., 2000), Probability of Switching (I-Index) (Guzmán et al., 2017), Burstiness (Goh and Barabási, 2008), and Complexity Factor (CF1-3) (Ghosh et al., 2017).

Table 1 shows the value of each metric for our genuine CSW dataset, as well as for these 3 synthetic CSW datasets. We can see that the LLM prompting-based dataset was superior in its similarity to the authentic CSW data. Using this method, we generated a corpus of 73,293 utterances covering over 20 English language pairs, including the 6 language pairs in the original dataset.[4]

| Metric | Genuine CSW | LLM CSW | Translation CSW | Co |
|---|---|---|---|---|
| CMI | 15.52 | 16.14 | 27.81 | |
| M-Index | 0.007 | 0.004 | 0.015 | |
| I-Index | 0.21 | 0.21 | 0.30 | |
| Burstiness | -0.07 | -0.04 | 0.03 | |
| CF1 | 6.38 | 5.82 | 17.13 | |
| CF2 | 19.77 | 19.03 | 31.11 | |
| CF3 | 18.34 | 17.61 | 30.05 | |

Table 1: Quantitative Description of the Genuine and Generated CSW Datasets Using Various CSW Metrics.

### 2.2.2 Step 2: Synthetic Error Generation

Several works have shown the effectiveness of rule-based error injection for GEC data generation. Many use the PIE-synthetic dataset (Awasthi et al., 2019), a perturbed version of the 1BW corpus (Chelba et al., 2013). For each sentence, the authors introduce between 0 and 4 errors of random type.

We extended this work by introducing a new subset of error types that are not only more common in ESL learners, but also are areas where the SOTA performance collapses when faced with CSW text: noun, pronoun, word order, determiner, and punctuation errors.

To increase the diversity of errors, we adopted a second style of error injection, Backtranslation (Stahlberg and Kumar, 2021). By swapping the source and target sentences of a monolingual dataset, we trained a GECToR-based system to induce errors in our synthetic CSW sentences. “‘

“‘latex

# 3 CSW GEC Systems

For our GEC system targeting CSW texts, we chose a GECToR model (Omelianchuk et al., 2020), with a RoBERTa-base foundation, due to its proven efficacy with limited training data and stronger performance on CSW texts compared to seq2seq models. We added a new CSW class to the error detection head, adding the ability to detect CSW tokens.

Following Tarnavskyi et al. (2022), we used a

[3] The full prompt can be seen in Appendix A.

[4] The LLM does not always generate the language pairs we ask for. However, these sentences are still included in the dataset categorised under their actual language pair.

[ILLEGIBLE]

Table 2: [ILLEGIBLE]

IMAGE NOT PROVIDED

[ILLEGIBLE]

Figure 2: [ILLEGIBLE]

3-stage training schedule. In the first, we used the same distilled 1BW corpus, and added all our synthetic CSW GEC data. In the second, we trained on a mixture of synthetic and genuine CSW GEC data. In the final stage, we fine-tuned only on the genuine CSW GEC dataset. “‘

“‘latex

# 4 Results and Analysis

## 4.1 Baseline Comparisons

[ILLEGIBLE]

## 4.2 Detailed Model Performance

[ILLEGIBLE]

[ILLEGIBLE]

## 4.3 Inference Tweaking and Error Thresholds

[ILLEGIBLE]

[ILLEGIBLE]

## 4.4 Synthetic Data Impact

[ILLEGIBLE]

[ILLEGIBLE]

# 5 Conclusion

[ILLEGIBLE]

[ILLEGIBLE]

# 6 Limitations

[ILLEGIBLE]

[ILLEGIBLE]

[ILLEGIBLE]
[ILLEGIBLE]
[ILLEGIBLE]

# References

# References

[1] [ILLEGIBLE]

[2] [ILLEGIBLE]

[3] [ILLEGIBLE]

[4] [ILLEGIBLE]

[5] [ILLEGIBLE]

[6] [ILLEGIBLE]

[7] [ILLEGIBLE]

[8] [ILLEGIBLE]

[9] [ILLEGIBLE]

[10] [ILLEGIBLE]

[11] [ILLEGIBLE]

[12] [ILLEGIBLE]

[13] [ILLEGIBLE]

[14] [ILLEGIBLE]

[15] [ILLEGIBLE]

[16] [ILLEGIBLE]

[17] [ILLEGIBLE]

[18] [ILLEGIBLE]

[19] [ILLEGIBLE]

[20] [ILLEGIBLE]

[21] [ILLEGIBLE]

[22] [ILLEGIBLE]

[23] [ILLEGIBLE]

[24] [ILLEGIBLE]

IMAGE NOT PROVIDED

[ILLEGIBLE]

Figure 3: [ILLEGIBLE]

[ILLEGIBLE]

Table 3: [ILLEGIBLE]

[25] [ILLEGIBLE]

[26] [ILLEGIBLE]

[27] [ILLEGIBLE]

[28] [ILLEGIBLE]

# A Example LLM Prompt

[ILLEGIBLE]

# B Error Type Analysis of SOTA

[ILLEGIBLE]

# C Training Data Schedule

[ILLEGIBLE]
[ILLEGIBLE]
[ILLEGIBLE]

# D Inference Hyperparameters

[ILLEGIBLE]
[ILLEGIBLE]

# E Error Type Analysis of Proposed Model

[ILLEGIBLE]

[ILLEGIBLE]

Table 4: [ILLEGIBLE]

[ILLEGIBLE]

Table 5: [ILLEGIBLE]

[ILLEGIBLE]

Table 6: [ILLEGIBLE]