Grammatical Error Correction for Code-Switched Sentences by Learners of English

Kelvin Wey Han Chan, Christopher Bryant, Li Nguyen, Andrew Caines, Zheng Yuan

University of Cambridge, Cambridge ALTA Institute for Automated Language Teaching and Assessment
WRITER
King's College London

Motivation

a. But the pay **a** little low . **[Monolingual English input]**
But the pay **is a** little low . **[GEC System output]**

b. But the 지불 **a** little low . **[Code-switching input]**
But the 지불 **a** little low . **[GEC System output]**

- While code-switching (CSW) benefits language learners, existing GEC systems are not designed for CSW, leading to errors in CSW sentences.
- This project aims to explore the task of GEC on CSW sentences, a novel and challenging domain that has not been explored before. This is done by generating synthetic CSW GEC datasets from existing GEC datasets.

CSW GEC Dataset Generation

Step	Sentence
1. Input monolingual	What if **human** use up all the **resource** in the world?
GEC data	What if **humans** use up all the **resources** in the world?
2. Select span	What if **humans** use up all the **resources** in the world?
3. Translate span	What if **humans** use up all the **resources** in the 世界?
4. Apply errors	What if **human** use up all the **resource** in the 世界?
5. Output CSW	What if **human** use up all the **resource** in the 世界?
GEC data	What if **humans** use up all the **resources** in the 世界?

Span Selection Methods

- **ratio-token:** randomly select tokens in the sentence until the ratio of tokens selected for CSW matches our reference CSW text.
- **cont-token:** randomly select a string of continuous tokens to match the ratio of CSW text.
- **rand-phrase:** randomly select phrases identified using the Berkeley Neural Parser ('benepar'). Linguistic research shows that CSW is usually based on a complete syntactic unit.
- **ratio-phrase:** select the phrase that has a token length closest to the reference number of tokens based on the CSW ratio.
- **overlap-phrase:** select the longest phrase that intersects with the least number of edit spans.
- **noun-token:** randomly select a single noun token within the sentence. Linguistic research shows that a majority of natural CSW only involves a single noun token

Experimental Setup

- GECToR is used as our baseline model. Compared to GECToR, the setup is modified in the following ways:
  - The training data from each stage is passed through our synthetic CSW data generation pipeline
  - XLM-RoBERTa is used as our pre-trained base model
- Evaluation was performed using ERRANT using the standard F0.5 metric

Evaluation Datasets

- **Lang-8 CSW Dataset:** A subset of the English Lang-8 Dataset with Chinese, Korean, and Japanese CSW components.
- **Human Re-annotated Dataset:** A subset (~200 sent. each) of the Chinese and Korean portions of the Lang-8 CSW Dataset re-annotated by bilingual Chinese-English and English-Korean speakers.

Results

Span Selection Method

- Linguistic insight is an important factor in synthetic CSW generation (i.e. rand-phrase and noun-token have the most consistent performance improvements over the baseline)

Method	Lang-8 CSW			Re-annotated CSW	
	ZH	KO	JA	ZH	KO
baseline	32.95	33.11	28.60	43.70	23.82
ratio-token	32.16	30.10	**28.76**	41.41	**23.92**
cont-token	**34.82**	28.82	27.53	**45.28**	21.06
rand-phrase	**33.99**	**34.42**	28.22	**46.28**	**25.35**
ratio-phrase	**34.40**	30.76	27.16	**45.27**	22.76
overlap-phrase	**33.95**	32.15	28.23	41.13	22.06
noun-token	**33.67**	**33.24**	**29.04**	**46.23**	**24.88**

Stage 1 Training Data

- Applying noun-token to all 3 training stages yields the best results

Training Stage			Lang-8 CSW			Re-annotated CSW	
1	2	3	ZH	KO	JA	ZH	KO
EN	EN	EN	37.16	33.46	30.02	45.74	24.23
EN	EN-CSW	EN-CSW	33.51	32.96	29.34	46.53	24.31
EN-CSW	EN-CSW	EN-CSW	**38.48**	**36.39**	**31.18**	**47.25**	**25.18**

Cross-Lingual Transferability

- Models trained on CSW text outperform monolingual models in all testing scenarios indicating that the CSW trained models generalize relatively well to other CSW languages.

Test Dataset		Training Dataset			
		EN-ZH	EN-KO	EN-JA	EN
Lang-8 CSW	ZH	**38.48**	37.87	36.81	37.16
	KO	36.08	**36.39**	36.25	33.46
	JA	31.22	**31.28**	31.18	30.02
Re-annotated CSW	ZH	**47.25**	46.68	46.61	45.74
	KO	22.98	**25.18**	23.19	24.23
BEA-19 Dev	EN	**52.42**	51.72	52.25	51.55

Conclusions

- We conducted the first study into developing GEC systems for CSW text
- We introduced a novel method of generating synthetic CSW GEC datasets using an existing GEC dataset and a translation model
- We introduced new CSW GEC datasets to evaluate our proposed models
- We found that randomly replacing a noun token in each sentence yielded the best improvement across the different methods, a conclusion supported by linguistic insight.
- Our findings show that models trained on one CSW language generalize relatively well to other languages and even improve performance on a benchmark monolingual dataset

The ALTA Institute is supported by Cambridge University Press & Assessment