Towards Minimal Supervision BERT-based Grammar Error Correction

Yiyuan Li, Antonios Anastasopoulos and Alan W Black

Language Technologies Institute

Carnegie Mellon University, Pittsburgh, PA

yiyuanli@andrew.cmu.edu {aanastas, awb}@cs.cmu.edu

Electrical & Computer ENGINEERING

Task

Grammartical Error Correction (GEC): Detect errors (misspelling, subject-verb agreement, determiner, etc.) in the sentences and correct them.

Input Plaing cards is bored, and exspensive .
Output Playing cards is boring and expensive .

Challenges

- Current grammatical error correction methods require large amount of annotated data, which may not be accessible in many languages.

Motivation

Utilizing grammatical information captured by unsupervised contextual model pre-trained on large corpora, like BERT [1] and extending to GEC in many languages with minimal supervision.

Dataset

- **Corpus:** The First Certificate in English (FCE), preprocessed to single error sentence-edit pairs (each edit and last edit).

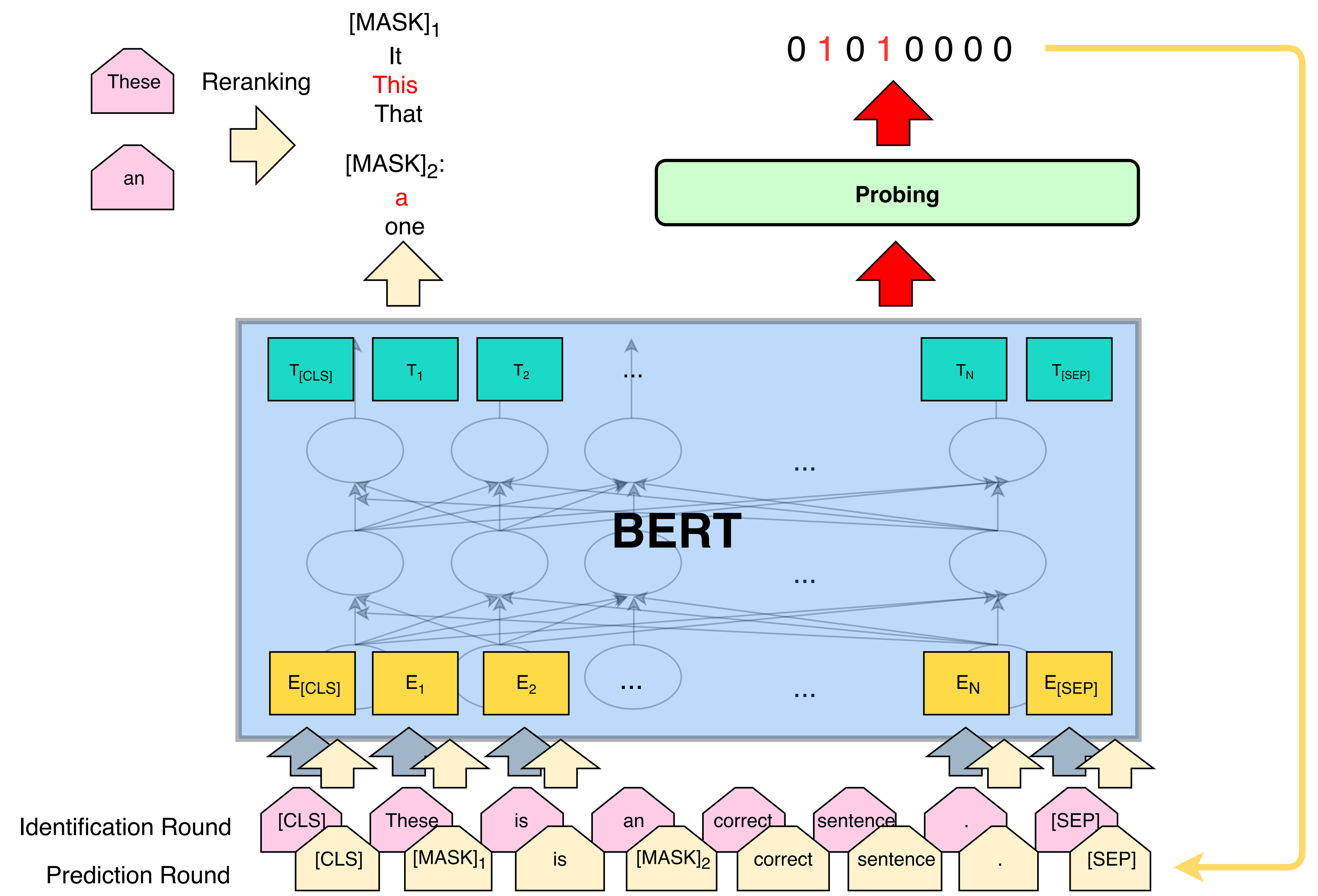
Proposed Method

Two stages:

- **Error Identification:**
  - BERT to detect.
  - Mask placement.
- **Mask Prediction:**
  - Predict token at masked position.
  - Rerank candidates.

[1] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL-HLT, 2019
[2] bert-base-multilingual-cased in https://github.com/huggingface/transformers/tree/master/transformers
[3] Kaneko, M., and Komachi, Multi-head multi-layer attention to deep language rep-resentations for grammatical error detection. Computaciony Sistemas, 2019

Model



Result

Table 1:Sentence-level evaluation.

Masking Strategy	each edit			last edit		
	P@1	R@1	$F_{0.5}$@1	P@1	R@1	$F_{0.5}$@1
# origin	0.632	0.853	0.667	0.592	0.824	0.627
# target	0.66	0.887	0.696	0.614	0.855	0.651
single	**0.763**	**0.931**	**0.790**	**0.767**	**0.920**	**0.794**

- Evaluate and predict on multilingual masked language model version of BERT [2] with different annotation schemes (# masks follow length of error span / correction / single).
- Sentence level evaluated by ERRANT (Table 1).
- Token level evaluated by performance@5 and performance@1 (Table 2).

Table 2:Token-level evaluation.

Masking Strategy	each edit		last edit	
	Acc@1	Acc@5	Acc@1	Acc@5
# origin	0.292	0.455	0.229	0.390
# target	0.313	0.484	0.247	0.405
single	**0.365**	**0.554**	**0.312**	**0.501**

- Multilingual BERT achieves precision over 0.7 without fine-tuning.
- Further potential improvement by reranking.

Error Analysis

Common BERT prediction errors, with the original error and the prediction highlighted.

Example #1: Redundant Edits	
Source	Of course there 's also a number 8 bus in front of the hotel , which is also suitable , but it leaves only every half an hour
Mask.	Of course there 's also a number 8 bus [MASK] in front of the hotel , which is also suitable , but it leaves only every half an hour
Target	Of course there 's also a number 8 bus , in front of the hotel , which is also suitable , but it leaves only every half an hour
Ours	Of course there 's also a number 8 bus stop in front of the hotel , which is also suitable , but it leaves only every half an hour
Example #2 : Synonyms	
Source	The aim of this report is to recomend you to visit the Fuerte de San Diego Museum
Mask.	The aim of this report is to [MASK] you to visit the Fuerte de San Diego Museum
Target	The aim of this report is to recommend you to visit the Fuerte de San Diego Museum
Ours	The aim of this report is to allow you to visit the Fuerte de San Diego Museum
Example #3 : Hallucination	
Source	Of course there 's also a bus number 8 , in front of the hotel , which is also suitable , but it leaves only every half an hour
Mask.	Of course there 's also a [MASK] [MASK] [MASK], in front of the hotel , which is also suitable , but it leaves only every half an hour
Target	Of course there 's also a number 8 bus , in front of the hotel , which is also suitable , but it leaves only every half an hour
Ours	Of course there 's also a small parking station , in front of the hotel , which is also suitable , but it leaves only every half an hour

Future Work

- **Error fertility:** Accurate mask placement.
- **Better span detection:** To leverage redundant edits.

Pre-trained Model	GED		
	P	R	$F_{0.5}$
BERT-base-uncased	0.480	0.359	0.450
BERT-multilingual-uncased	0.464	0.319	0.426
Kaneko and Komachi 2019 [3]	0.698	0.374	0.595

Table 3:Performance of fine-tuning BERT on FCE for grammatical error detection (GED)

- **Masking and fluency measure:** To limit unwanted freedom in prediction, and set up criterion for iterative editing.

Conclusion

- Pre-trained BERT can achieve more than 0.7 precision in single-error grammatical error correction without fine-tuning, and could be potentially improved by re-ranking.
- Advanced masking and fluency measure are needed to leverage information lost by masking and setting up ending criterion in editing.