Bidirectional Transformer Reranker for Grammatical Error Correction

Ying Zhang **Hidetaka Kamigaito** **Manabu Okumura**

{ zhang, oku } @ lr.pi.titech.ac.jp kamigaito.h @ is.naist.jp

Homepage:

Ying Zhang's website

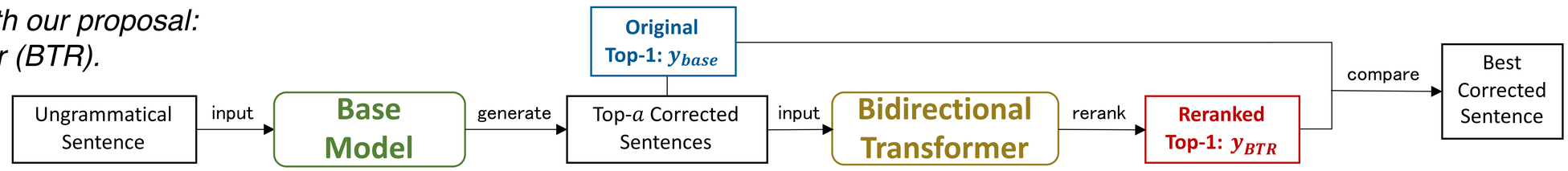
Paper:

東京工業大学 Tokyo Institute of Technology

国立大学法人 奈良先端科学技術大学院大学 NARA INSTITUTE of SCIENCE and TECHNOLOGY

理化学研究所 RIKEN

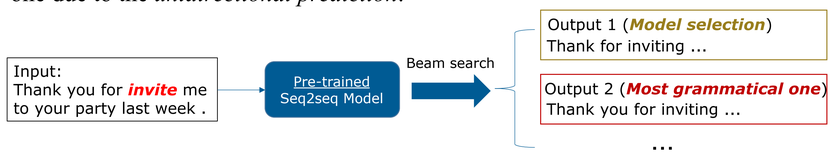
Reranking corrected sentences with our proposal: Bidirectional Transformer Reranker (BTR).



Motivation

Grammatical error correction: a sequence-to-sequence (seq2seq) task, which requires a model to correct an ungrammatical sentence.

While a fully pre-trained seq2seq model can generate several high-quality grammatical sentences using beam search and even achieve state-of-the-art results, there may still be a gap between the selected hypothesis and the most grammatical one due to the *unidirectional prediction*.



Therefore, it is potential for improving the performance of pre-trained seq2seq models by utilizing the entire representations of the prediction, i.e., the *bidirectional representations*.

Experiments

Datasets

Dataset	Usage	Lang	# of data (pairs)
Realnewslike	pre-train	EN	148,566,392
cLang-8	train	EN	2,372,119
CoNLL-13	valid	EN	1,381
CoNLL-14	test	EN	1,312
BEA	test	EN	4,477
JFLEG	test	EN	747

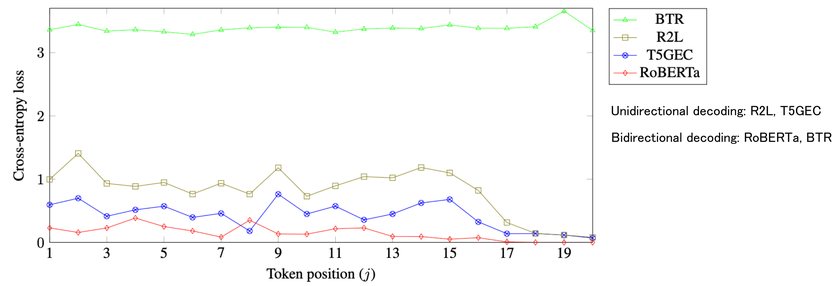
Table 1: Dataset sizes.

Main Results

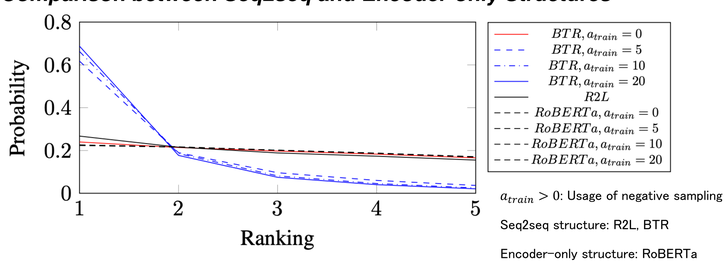
Model	CoNLL-13 ($F_{0.5}$ ↑)	CoNLL-14 ($F_{0.5}$ ↑)	BEA-test ($F_{0.5}$ ↑)	JFLEG (GLEU↑)
Oracle	55.11	67.87	-	61.13
T5GEC*	-	65.13	69.38	-
T5GEC	49.36	65.11	70.51	59.04
R2L	50.02	64.92	**71.42**	58.93
w/o L2R	49.19	64.54	69.76	58.69
RoBERTa ($\lambda = 0.1$) w/o a_{train}	49.35	65.04	70.55	59.17
w/o a_{train}, λ	46.48	60.90	64.05	57.49
BTR($\lambda = 0.4$)	**50.22**	**65.47**	71.27	59.17
w/o λ	49.13	63.82	69.93	**59.52**
w/o a_{train}, λ	45.34	59.48	63.60	57.62

Table 2: Results for the models on each dataset with candidates from T5GEC.

Comparison between Unidirectional and Bidirectional Decoders

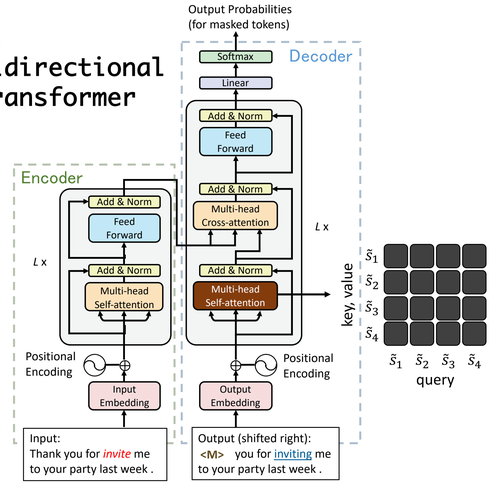


Comparison between Seq2seq and Encoder-only Structures



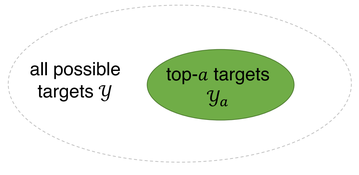
Methods

1. Bidirectional Transformer



Usage	Inputs x	Masked Targets $y_{\setminus\kappa}$
Self-supervised learning for pre-training	Thank you <X> me to your party <Y> week .	<X> for <M> *you* last <Z>
Supervised learning for fine-tuning	Thank you for *invite* me to your party last week .	Thank you *so* <M> me to your party <M> week .

Table 3: Examples of data pairs.



2. Target Sentence Probability

$$\log p(\boldsymbol{y}|\boldsymbol{x};\theta) \approx \mathrm{PLL}(\boldsymbol{y}|\boldsymbol{x};\theta) = \sum_{j=1,\kappa=\{j\}}^{|\boldsymbol{y}|} \log p(y_j|\boldsymbol{x},\boldsymbol{y}_{\setminus\kappa};\theta)$$

3. Score Target Sentence

$$f(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp\left(\mathrm{PLL}(\boldsymbol{y}|\boldsymbol{x};\theta)/|\boldsymbol{y}|\right)}{\sum_{\boldsymbol{y}'\in\mathcal{Y}_a}\exp\left(\mathrm{PLL}(\boldsymbol{y}'|\boldsymbol{x};\theta)/|\boldsymbol{y}'|\right)}$$

4. Accept Target sentence

$$\text{Final result} = \begin{cases} \boldsymbol{y}_{BTR} \text{ if } f(\boldsymbol{y}_{BTR}|\boldsymbol{x}) - f(\boldsymbol{y}_{base}|\boldsymbol{x}) > \lambda, \\ \boldsymbol{y}_{base} \text{ if } f(\boldsymbol{y}_{BTR}|\boldsymbol{x}) - f(\boldsymbol{y}_{base}|\boldsymbol{x}) \le \lambda \end{cases}$$

4. Objective Function

$y_{gold} \in \mathcal{Y}$: gold correction for the given x.

For $\boldsymbol{y} \in \mathcal{Y}' \cup \{\boldsymbol{y}_{gold}\}$, we follow the setting of BERT to *randomly mask* 15% of $\boldsymbol{y}$ as $\boldsymbol{y}_{\setminus\kappa}$, where κ denotes the set of masked positions. The distribution of the masked tokens satisfies the 8:1:1 masking strategy.

Given the masked target $\boldsymbol{y}_{\setminus\kappa}$, the model parameter θ of the BTR is optimized by:

$$\log p(\boldsymbol{y}_\kappa|\boldsymbol{x},\boldsymbol{y}_{\setminus\kappa};\theta) \approx \sum_{j\in\kappa}[\mathbb{1}\underbrace{\log p(y_j|\boldsymbol{x},\boldsymbol{y}_{\setminus\kappa};\theta)}_{\text{maxmize likelihood}} + \underbrace{(1-\mathbb{1})\log(1-p(y_j|\boldsymbol{x},\boldsymbol{y}_{\setminus\kappa};\theta))}_{\text{minimize unlikelihood}}]$$

Where $\mathbb{1}$ is the indicator function, defined as follows:

$$\mathbb{1} := \begin{cases} 1 & \text{if } \boldsymbol{y} = \boldsymbol{y}_{gold}, \\ 0 & \text{if } \boldsymbol{y} \neq \boldsymbol{y}_{gold} \end{cases}$$