

Paper review

LM-Critic: Language Models for Unsupervised Grammatical Error Correction (EMNLP 2021)

Presentation: **Jeiyoon Park**
6th Generation, TAVE

Outline

1. Contribution
2. Method
3. Experiments
4. Conclusion

Outline

1. Contribution
2. Method
3. Experiments
4. Conclusion

Detour: Grammatical Error Correction

1. Grammatical Error Correction (GEC)

24. 다음 글의 밑줄 친 부분 중, 어법상 틀린 것은? [3점]

In some communities, music and performance have successfully transformed whole neighborhoods as ① profoundly as The Guggenheim Museum did in Bilbao. In Salvador, Brazil, musician Carlinhos Brown established several music and culture centers in formerly dangerous neighborhoods. In Candeal, ② where Brown was born, local kids were encouraged to join drum groups, sing, and stage performances. The kids, energized by these activities, ③ began to turn away from dealing drugs. Being a young criminal was no longer their only life option. Being musicians and playing together in a group looked like more fun and was more ④ satisfying. Little by little, the crime rate dropped in those neighborhoods; the hope returned. In another slum area, possibly inspired by Brown's example, a culture center began to encourage the local kids to stage musical events, some of ⑤ them dramatized the tragedy that they were still recovering from.




Great Writing, Simplified

Compose bold, clear, mistake-free writing with Grammarly's AI-powered writing assistant.

[Add to Chrome](#) It's free


★★★★★ 34,000+ Chrome store reviews
20 million people use Grammarly to improve their writing



Hi Jen,

I hope your well. Can we catch up today? I'd really apprec you're intation for tomori / love it, if you could ouuue-check the sales numbers with me. There's a coffee in it for you!

CORRECTNESS: SPELLING



[전체](#) [이미지](#) [도서](#) [뉴스](#) [동영상](#) [더보기](#) [도구](#)

검색결과 약 2,050,000,000개 (0.51초)

수정된 검색어에 대한 결과: **terrestrial**
다음 검색어로 대신 검색: **terrestrial**

Detour: Grammatical Error Correction

1. Grammatical Error Correction (GEC)

- 1) GEC is the task of fixing grammatical errors in text, such as typos, tense and article mistakes
- 2) Training a model for GEC requires a set of labeled *(ungrammatical / grammatical)* sentence pairs, which are expensive to obtain

(1) She like cats.

(2) Nothing is absolute right or wrong. (absolutely)

(3) One option to moving toward both biodiversity and terrestrial food supply goals are to produce greater yield from less land



(1) She **likes** cats.

(2) Nothing is **absolutely** right or wrong.

(3) One option **for** moving toward both biodiversity and **terrestrial** food supply goals **is** to produce greater **yields** from less land

Detour: Grammatical Error Correction

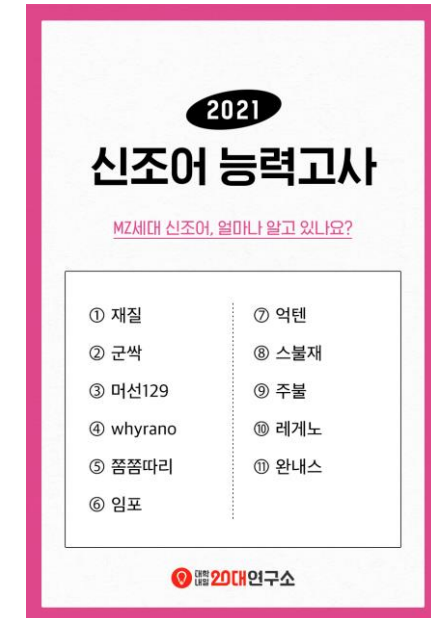
2. Challenges

1) Due to the **unrestricted mutability of language**, it is hard to design a model that is capable of correcting all possible errors made by non-native learners, especially when error patterns in new text are not observed in training data.

2) Unlike machine translation, **a large amount of annotated ungrammatical texts and their corrected counterparts** are not available.

3) The **artificially generated data** cannot precisely capture the error distribution in real erroneous data.

e.g.) We don't use “*a am I boy*” (“*I am a boy*”)



Detour: Grammatical Error Correction

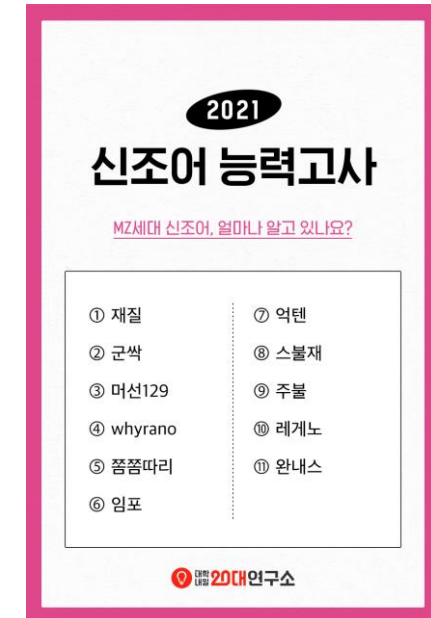
2. Challenges

1) Due to the **unrestricted mutability of language**, it is hard to design a model that is capable of correcting all possible errors made by non-native learners, especially when error patterns in new text are not observed in training data.

✓ 2) Unlike machine translation, **a large amount of annotated ungrammatical texts and their corrected counterparts** are not available.

✓ 3) The **artificially generated data** cannot precisely capture the error distribution in real erroneous data.

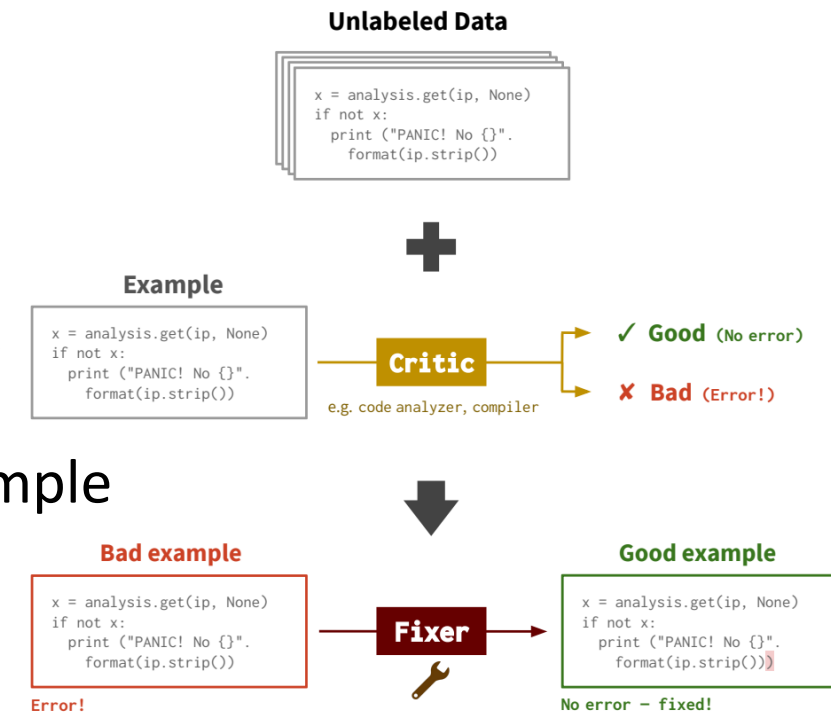
e.g.) We don't use "a am I boy" ("I am a boy")



Contribution

1. Break-It-Fix-It (BIFI) Framework

- They apply BIFI (Yasunaga et al., ICML 2021) to GEC task
- BIFI consists of (1) Breaker, (2) Fixer, and (3) Critic
- **Breaker**: Generate realistic bad example from good example
- **Fixer**: Converts a bad example into a good one
- **Critic**: Check fixer's output on real bad inputs



2. LM-Critic

- However, a “perfect” critic that returns whether an example is good or bad **doesn't exist**
- They leverage a pretrained language model (LM) in defining an **LM-Critic**, which judge a sentence to be grammatical if the **LM assigns it a higher probability than its local perturbations**

Outline

1. Contribution
- 2. Method**
3. Experiments
4. Conclusion

Problem Setup

- Notation

x_{bad} : Ungrammatical sentence

x_{good} : Grammatical version of x_{bad}

f : A GEC model (a.k.a. **Fixer**)

$D_{\text{pair}} = \{(x_{\text{bad}}^{(i)}, x_{\text{good}}^{(i)})\}$: A paired dataset

labeled: the pairs are human-annotated

unlabeled: A set of raw sentences $D_{\text{unlabel}} = \{x^{(i)}\}$

$$\text{critic } c: c(x) = \begin{cases} 1 & \text{if } x \text{ is good} \\ 0 & \text{if } x \text{ is bad} \end{cases}$$

Problem Setup

- Notation

x_{bad} : Ungrammatical sentence

x_{good} : Grammatical version of x_{bad}

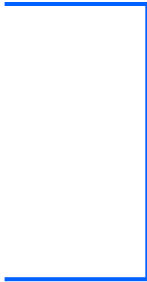
f : A GEC model (a.k.a. **Fixer**)

$D_{\text{pair}} = \{(x_{\text{bad}}^{(i)}, x_{\text{good}}^{(i)})\}$: A paired dataset

labeled: the pairs are human-annotated

unlabeled: A set of raw sentences $D_{\text{unlabel}} = \{x^{(i)}\}$

critic c : $c(x) = \begin{cases} 1 & \text{if } x \text{ is good} \\ 0 & \text{if } x \text{ is bad} \end{cases}$



Given D_{unlabel} and **LM**, which returns a probability distribution $p(x)$ over sentence x , we can **define the critic** and **use that to the fixer**

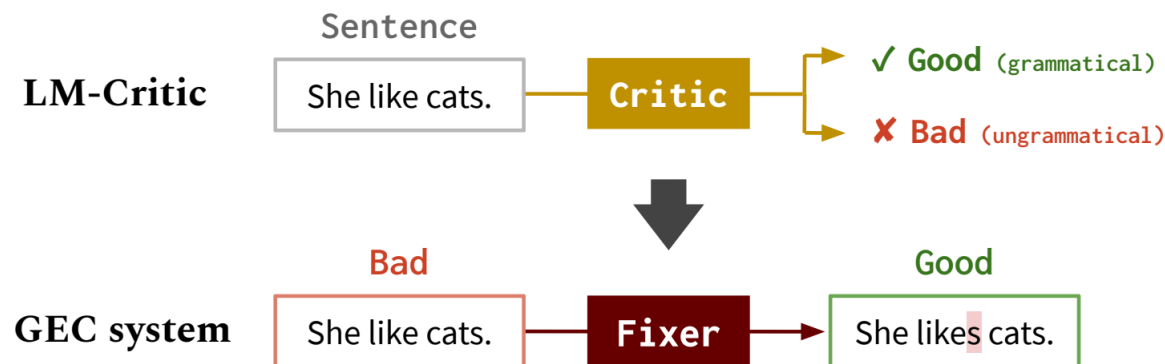
Method: LM-Critic

- Criterion

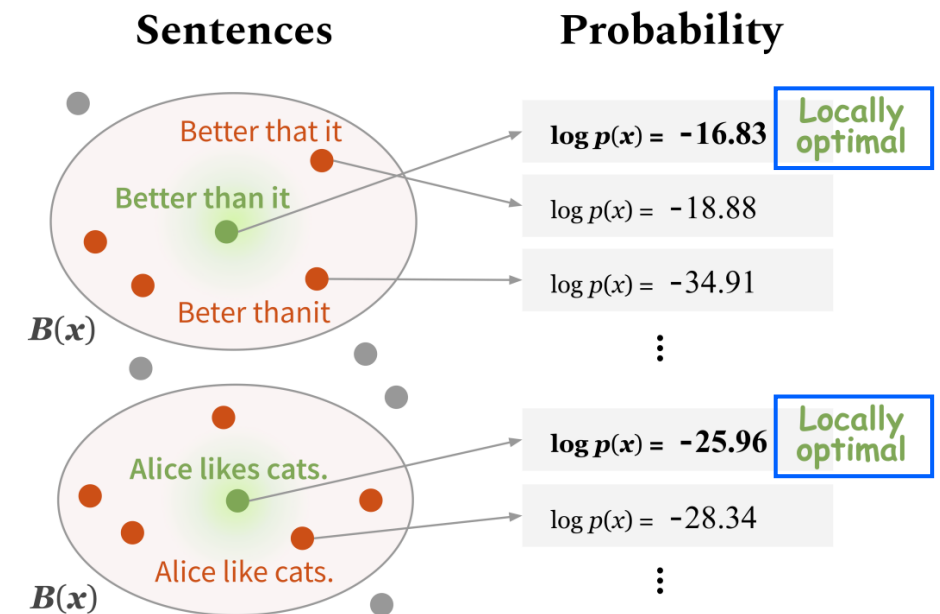
It deems a sentence to be **good** if it has the highest probability within its local neighborhood (local optimum criterion)

- Implementation

(1) A pretrained LM, and (2) perturbation function



(b) Idea behind LM-Critic: Local optimum criterion



Method: LM-Critic

1. Local Optimum Criterion of Grammaticality

(1) Starting point

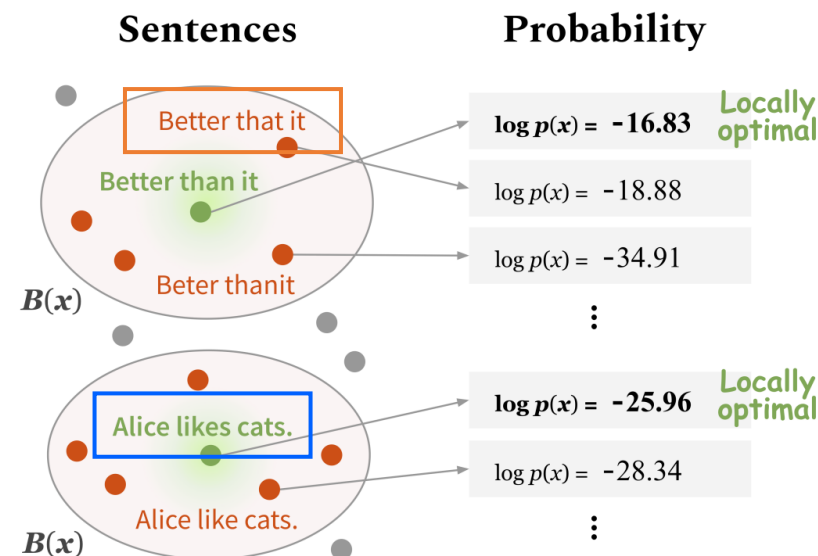
- Find a threshold (δ) for the absolute probability and let the critic be:

$$\text{AbsThr-Critic}(x) = \begin{cases} 1 & \text{if } p(x) > \delta \\ 0 & \text{otherwise.} \end{cases}$$

- However, it doesn't work in practice

e.g.) $\log p(\text{"Alice likes cats"}) < \log p(\text{"Better that it"})$

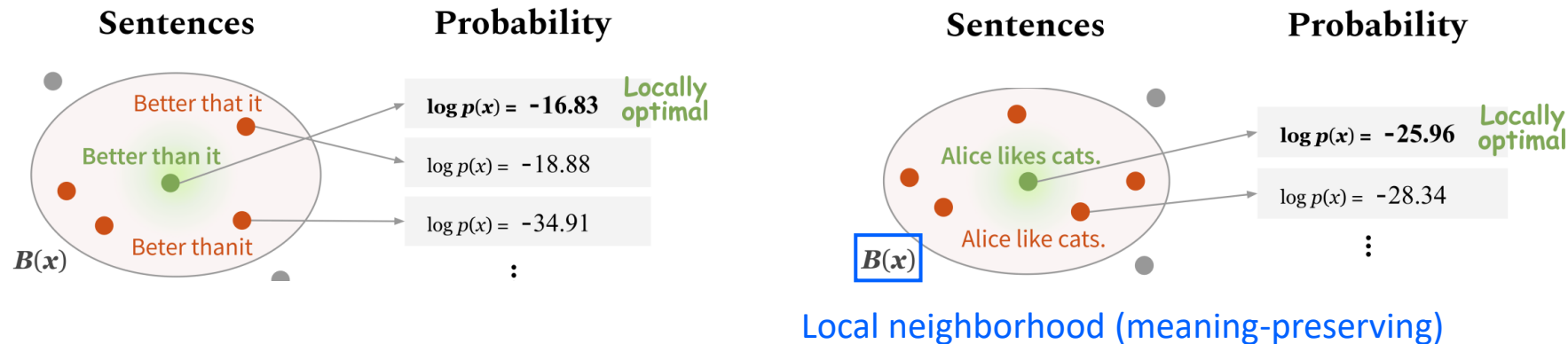
(b) Idea behind LM-Critic: Local optimum criterion



Method: LM-Critic

1. Local Optimum Criterion of Grammaticality

(2) LM-Critic compares sentences with the same intended meaning



(3) Local optimum criterion of grammaticality:

$$x \text{ is grammatical iff } x = \operatorname{argmax}_{x' \in B(x)} p(x').$$

Method: LM-Critic

2. Implementation of LM-Critic

- Obtaining ground-truth local neighborhood $B(x)$ is difficult \rightarrow samples $\hat{B}(x)$

$$\text{LM-Critic}(x) = \begin{cases} 1 & \text{if } x = \underset{x' \in \hat{B}(x)}{\operatorname{argmax}} p(x') \\ 0 & \text{otherwise.} \end{cases}$$

- There are three decisions for implementing LM-Critic:

- (1) Choice of a pretrained LM
- (2) Perturbation function b
- (3) Sampling method of perturbations

Method: LM-Critic

2. Implementation of LM-Critic

(1) Choice of a pretrained LM

- **GPT2 (#: 117M)**
 - **GPT2-medium (#: 345M)**
 - **GPT2-large (#: 774M)**
 - **GPT2-xl (#: 1.6B)**
- The LMs were trained on a large set of web text (40GB)

Method: LM-Critic

2. Implementation of LM-Critic

(2) Perturbation function b

- **ED1**: Edit-distance one perturbation in the character space (e.g. insert, delete, replace, and swap two adjacent characters)
- **ED1 + Word-level heuristics (all)**: ED1 can't fully cover word-level errors. It includes heuristics for word-level perturbations based on its dictionary
- **ED1 + Word-level heuristics**: It removes some heuristics that alter the meaning of the original sentence

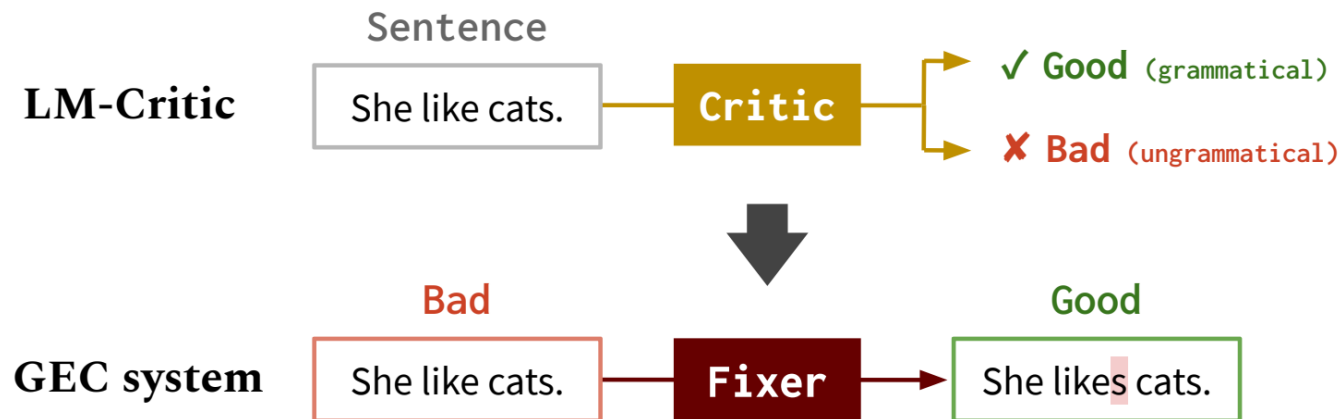
e.g.) deleting / inserting “not”

Method: LM-Critic

2. Implementation of LM-Critic

(3) Sampling method of perturbations

- Random sampling with sizes of 100, 200, and 400
- They obtain samples from $b(x)$ to be $\hat{B}(x)$



Method: LM-Critic

3. Empirical Analysis

- Simple check: To make sure that LM's probability score correlates with grammaticality

(1) Evaluation data

- They prepare a simple evaluation data consisting of (x_{bad}, x_{good})
- They combine the dev sets of multiple GEC benchmarks, including *GMEG-wiki*, *GMEG-yahoo*, *BEA-2019*
- #: about 600

Method: LM-Critic

3. Empirical Analysis

(2) Analysis of LM probability

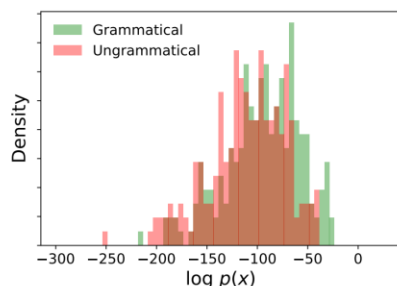


Figure 2: Probability of grammatical (green) and ungrammatical (red) sentences, computed by a pretrained LM (GPT2).

Pretrained LM	How often $p(x_{\text{bad}}) < p(x_{\text{good}})$?
GPT2	94.7%
GPT2-medium	95.0%
GPT2-large	95.9%
GPT2-xl	96.0%

Table 1: How well sentence probability returned by pretrained LMs correlates with grammaticality empirically.

- Remaining pairs (5.3%)

Examples of $p(x_{\text{bad}}) > p(x_{\text{good}})$

(Comma)

x_{bad} : The video was filmed on January 22 and is set to premiere on February 22.

x_{good} : The video was filmed on January 22, and is set to premiere on February 22.

(Quotation)

x_{bad} : Uprising is a 1980 roots reggae album by Bob Marley & The Wailers.

x_{good} : "Uprising" is a 1980 roots reggae album by Bob Marley & The Wailers.

(British spelling)

x_{bad} : The blast could be heard across the whole city centre.

x_{good} : The blast could be heard across the whole city center.

Examples of $p(x') > p(x_{\text{good}})$, $x' \in \hat{B}(x_{\text{good}})$

(Singular/plural)

x' : They are affiliated to either the state boards or to national education boards.

x_{good} : They are affiliated to either the state board or to national education boards.

(Tense)

x' : As well as touring Europe, they tour with such acts as Green Day.

x_{good} : As well as touring Europe, they toured with such acts as Green Day.

Table 3: Failure cases of LM-Critic. (Top) GPT2 assigns a higher probability to bad sentences. (Bottom) our neighborhood function ("ED1 + word") includes sentences with a higher LM probability than the original good sentence.

Method: LM-Critic

3. Empirical Analysis

(3) Performance of LM-Critic (using evaluation set)

Perturbation	Recognize “Good”			Recognize “Bad”		
	P	R	F _{0.5}	P	R	F _{0.5}
ED1	58.7	90.1	63.1	78.8	36.8	64.2
ED1 + word(all)	69.7	10.2	32.2	51.5	95.5	56.7
ED1 + word	68.4	75.5	69.7	72.7	65.1	71.1

Sample size	Recognize “Good”			Recognize “Bad”		
	P	R	F _{0.5}	P	R	F _{0.5}
100	68.4	75.5	69.7	72.7	65.1	71.1
200	71.3	71.5	71.4	71.4	71.3	71.4
400	72.6	68.7	71.8	70.3	74.0	71.0

Pretrained LM	Recognize “Good”	Recognize “Bad”
	F _{0.5}	F _{0.5}
GPT2	69.7	71.1
GPT2-medium	69.9	71.0
GPT2-large	70.3	71.3
GPT2-xl	69.9	71.0



Table 2: **Performance of LM-Critic**, when using different choices of a perturbation function, sample size, and pretrained LM described in §3.2. **(Top)** We set the LM to be GPT2 and the perturbation sample size to be 100, and vary the perturbation function b . “ED1 + word” achieves the best $F_{0.5}$. Henceforth, we use this perturbation function. **(Middle)** We set the LM to be GPT2 and vary the perturbation sample size. Increasing the sampling size improves the performance slightly. **(Bottom)** We vary the LM. Increasing the LM size makes slight or no improvement in $F_{0.5}$ on the dataset we used.

Method: LM-Critic

4. Learning GEC with LM-Critic

- Initial fixer f_0 is trained on synthetic data (unsupervised setting) or labeled data (supervised setting)

- (1) Apply the fixer f to the bad example D_{bad} (by human)

- (2) They train a *breaker* b on resulting paired data

- (3) They apply the breaker to the good example D_{good}

- (4) They finally train the fixer on the newly-generated paired data in (1) and (3)

- This cycle can be iterated to improve both fixer and breaker

Outline

1. Contribution
2. Method
- 3. Experiments**
4. Conclusion

Experiments

1. Evaluation data: GEC benchmarks

- CoNLL-2014
- BEA-2019
- GMEG-yahoo
- GMEG-wiki

Experiments

2. Unsupervised setting

- Training data: **One-billion-word corpus** (Chelba et al., 2013)
- Synthetic paired data by heuristically corrupting sentences
- No labeled training data
- #: 9M (pairs) → Training baseline fixer (encoder-decoder Transformer)

GEC system	CoNLL-2014 (test)			BEA-2019 (dev)			GMEG-wiki (test)			GMEG-yahoo (test)		
	P	R	F _{0.5}	P	R	F _{0.5}	P	R	F _{0.5}	P	R	F _{0.5}
Transformer	59.2	29.2	49.1	44.2	17.9	34.1	52.1	26.5	43.7	44.4	36.9	42.7
+ BIFI with no critic	58.2	29.9	48.9	43.8	18.7	34.5	53.5	27.4	44.9	45.1	38.5	43.6
+ BIFI (ours)	64.4	35.6	55.5	51.6	24.7	42.4	57.9	33.6	50.6	53.7	47.1	52.2

Table 4: GEC results in the **unsupervised setting** (§4.2.2). “Transformers” is trained on synthetic paired data as in [Awasthi et al. \(2019\)](#). If we train it on more realistic paired data generated by BIFI (bottom row), it achieves improved results.

Experiments

3. Supervised setting

- Training data: CoNLL-2014, and BEA-2019
- **GECToR** (Omelianchuk et al., 2020) as baseline fixer
- GECToR is first trained on the synthetic paired data and then trained on the labeled data

GEC system	Ens.	CoNLL-2014 (test)			BEA-2019 (test)		
		P	R	F _{0.5}	P	R	F _{0.5}
GPT3 (175B) with prompting		62.4	25.0	48.0	50.8	38.2	47.6
Zhao et al. (2019)		67.7	40.6	59.8	-	-	-
Awasthi et al. (2019)		66.1	43.0	59.7	-	-	-
Kiyono et al. (2019)		67.9	44.1	61.3	65.5	59.4	64.2
Zhao et al. (2019)	✓	74.1	36.3	61.3	-	-	-
Awasthi et al. (2019)	✓	68.3	43.2	61.2	-	-	-
Grundkiewicz et al. (2019)	✓	-	-	64.2	72.3	60.1	69.5
Kiyono et al. (2019)	✓	72.4	46.1	65.0	74.7	56.7	70.2
Kantor et al. (2019)	✓	-	-	-	78.3	58.0	73.2
GECToR (Omelianchuk et al., 2020)		77.5	40.1	65.3	79.2	53.9	72.4
GECToR (our base)		77.5	40.1	65.3	79.2	53.9	72.4
+ BIFI (ours)		78.0	40.6	65.8	79.4	55.0	72.9

Table 5: GEC results in the **supervised setting** with labeled data available (§4.2.3). “Ens.” indicates an ensemble system.

Outline

1. Contribution
2. Method
3. Experiments
- 4. Conclusion**

Conclusion

1. Analysis

(a) Pairs generated by synthetic corruption	
x_{bad} :	We look forward the to better treatments in the future.
x_{good} :	We look forward to better treatments in the future.
x_{bad} :	The president-elect stayed away so as not to foregin matters until Bush.
x_{good} :	The president-elect stayed away so as not to complicate matters for Bush.
(b) Pairs generated by BIFI without LM-Critic	
x_{bad} :	If anyone is interested, here's the kink.
x_{good} :	If anyone is interested, here's the kinks.
x_{bad} :	If you can't find a match yourself, horse trader will helps.
x_{good} :	If you can't find a match yourself, horse traders will help.
(c) Pairs generated by BIFI with LM-Critic (Ours)	
x_{bad} :	First Light is a award-winning novel by Sunil Gangopadhyay.
x_{good} :	First Light is an award-winning novel by Sunil Gangopadhyay.
x_{bad} :	Except latter, the rivers are in underground tubes and not visible.
x_{good} :	Except for the latter, the rivers are in underground tubes and not visible.

Table 6: Examples of paired data generated by (a) synthetic corruption, (b) BIFI without critic, and (c) BIFI with LM-Critic. (a) tends to deviate from the type of grammatical errors humans make. (b) tends to have pairs where x_{good} is broken (*e.g.*, the first pair) or x_{bad} is already grammatical, as pairs are not verified by a critic. (c) is the most realistic.

(Input) The system is designed to use amplitude comparison for height finding.	
(Baseline)	The system is designed to use amplitude comparison for height find .
(BIFI)	The system is designed to use amplitude comparison for height finding.
(Input) Lugu Lake, set in the subalpine zone in Hengduan is a landscape of pine-covered ecoregion.	
(Baseline)	Lugu Lake, set in the subalpine zone in Hengduan, is their landscape of pine-covered ecoregion.
(BIFI)	Lugu Lake, set in the subalpine zone in Hengduan, is a landscape of pine-covered ecoregion.

Table 7: Examples where the baseline fixer trained with synthetic data fails but BIFI succeeds. The baseline tends to make unnecessary edits (*e.g.*, changing verb inflection or articles, due to heuristics used when generating synthetic data).

Conclusion

2. Any drawbacks?: Assumptions

(1) They excludes an ungrammatical sentence which may have no correction

e.g.) “asdfghgfdsa”

(2) They also didn’t consider multiple corrections

e.g.) “The cat sleep” → “The cat sleep^s”? or “The cat ^{slept}”?



Thank you

<https://jeiyoong.github.io/>