Converge to the Truth: Factual Error Correction via Iterative Constrained Editing

Jiangjie Chen[*1], Rui Xu[*1], Wenxuan Zeng[2], Changzhi Sun[3], Lei Li[4], Yanghua Xiao[1]

FUDAN UNIVERSITY
University of Electronic Science and Technology of China
ByteDance
UC SANTA BARBARA

Factual Errors in Texts

Academic writing

Advances in neural information processing systems, 2019, 30.

Advances in neural information processing systems, 2017, 30.

Journalism

James Cameron directed Thor 2, which was released in 2022.

James Cameron directed Avatar 2, which was released in 2022.

Online content/ AIGC

Socrates wrote the Ethics and the Republic.

Platos wrote the Ethics and the Republic.

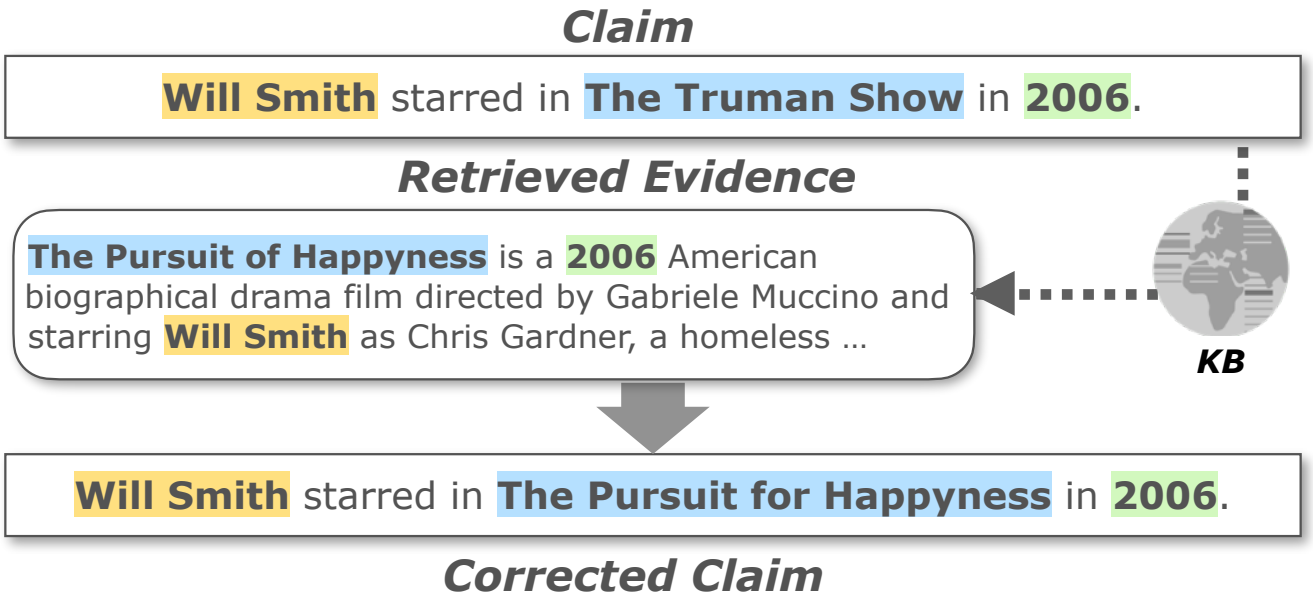
Correcting Textual Errors

- ✍️ Grammatical error correction
  - Error forms are easy to summarize.
  - Massive data for supervised models.
  - No external knowledge is required.

> Nothing is [**absolute** -> **absolutely**] right or wrong.
>
> There are [**any** -> **some**] apples on the table.
>
> Don't drink [**or** -> **and**] drive.

- How about factual error correction?

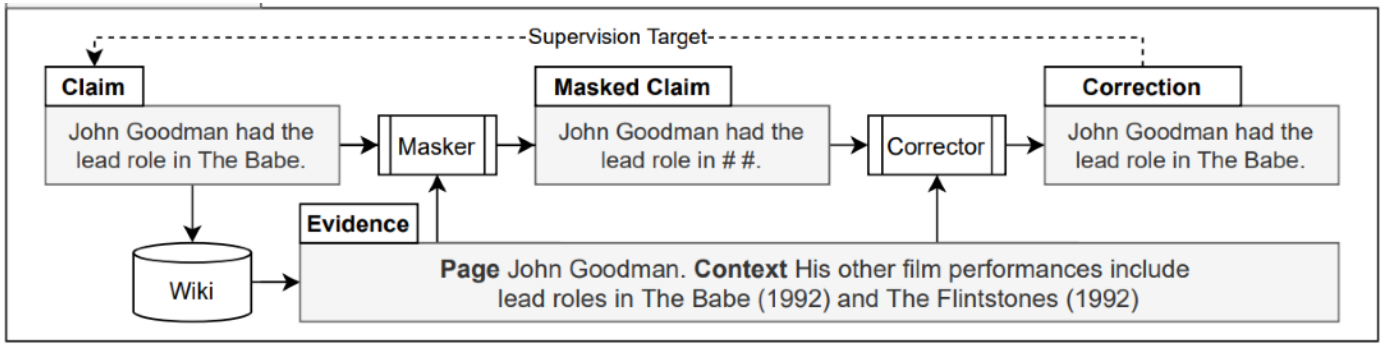
Evidence-based Factual Error Correction



- ! **BUT**: Lack fine-grained annotations and high-quality datasets, which are costly to build.
- Most datasets are synthetically built.

No Supervised FEC Data? Fact Verification Helps!

💡 **Previous methods**: one-pass mask-then-correct generation



! Challenges for Distant Supervised Methods

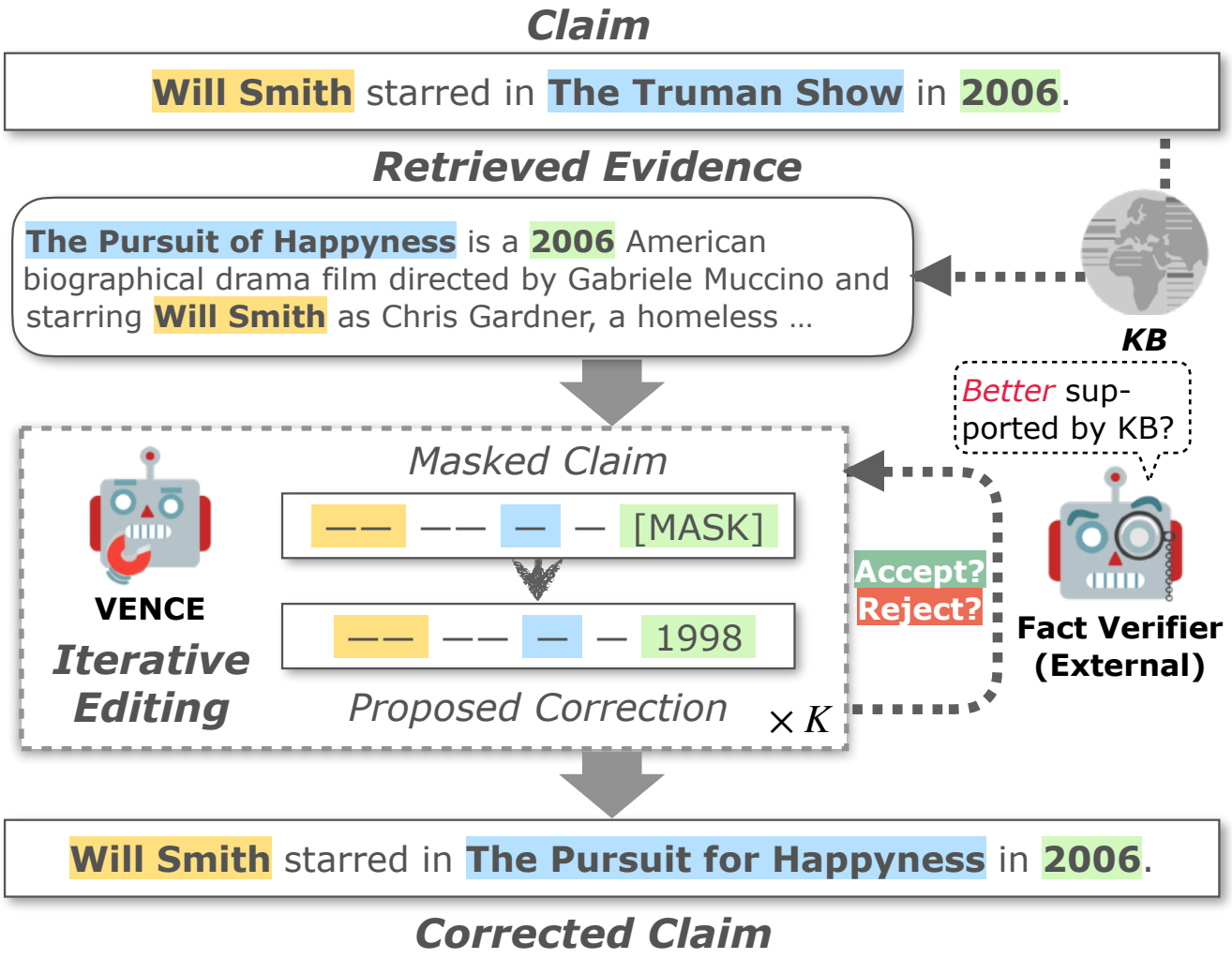
- **Over-erasure**: there may be too many masked spans that overwrite the intended meaning of the claim.
  - "New York has 18 million population." => "New York is in the US."
- **Missing validation**: They did not explicitly consider validating the veracity of the corrected claim, which may still be nonfactual.

Motivations

- **Over-erasure**: Correct errors via iterative editing
  - Break the correction process into unit-level (token/ entity) to revise more choices.

- **Missing validation**: Bridge Fact Verification with Factual Error Correction
  - FV offers control and guidance to the correction in each editing iteration.
  - Resources for FV are significantly richer than FEC.

VENCE: VErificatioN-guided Constrained Editing



Energy Functions in $\pi(x)$

Desired properties of the target texts

$$\mathscr{E}(x) = \mathscr{E}_{\mathtt{LM}}(x) + \mathscr{E}_{\mathtt{V}}(x) + \mathscr{E}_{\mathtt{H}}(x)$$

Fluency:
- Language Modeling

$$\mathscr{E}_{\mathtt{LM}}(x) = -\sum_i \log P_{\mathtt{MLM}}(w_i \,|\, x_{-i})$$

Truthfulness:
- Fact Verification

$$\mathscr{E}_{\mathtt{V}}(x) = -\log P_{\mathtt{V}}(\text{Supported} \,|\, x, E)$$

Minimal-edits:
- Hamming Distance

$$\mathscr{E}_{\mathtt{H}}(x) = \text{HammingDistance}(x, x^0)$$

Constrained Text Editing via Metropolis-Hastings Sampling (Metropolis et al. 1953)

Stationary distribution

Where we want the sampling to converge

$$\pi(x) = \frac{e^{-\mathscr{E}(x)}}{Z}$$

Transition distribution

In the Markov chain, taking the action a to edit position m

$$g(x' \mid x) = P_1(m \mid x)P_2(a)P_3(x' \mid x_{-m}, a)$$

Acceptance Ratio

Decides the acceptance of each proposal

$$A(x' \mid x) = \min\{1, \frac{\pi(x')g(x \mid x')}{\pi(x)g(x' \mid x)}\}$$

$$= \min\{1, \frac{e^{-\mathscr{E}(x')}g(x \mid x')}{e^{-\mathscr{E}(x)}g(x' \mid x)}\}$$

Workflow

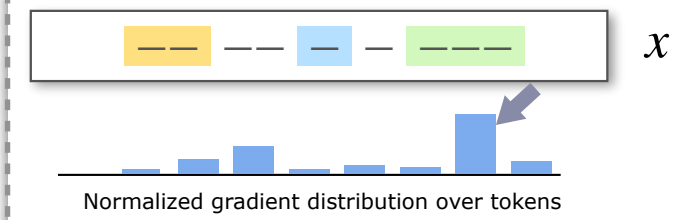
Input: x

Workflow

Input: x

↓

Sample a position to edit with a verifier



Editing Position Proposal: $P_1(m \mid x)$

- Sample a position based on Verifier (P_V)'s **normalized gradient distribution**.

- **Multi-token Entity Masking**
  - ✓ Will Smith starred in [MASK] in 2006.
  - – ~~Will Smith starred in The [MASK] Show in 2006.~~

Workflow

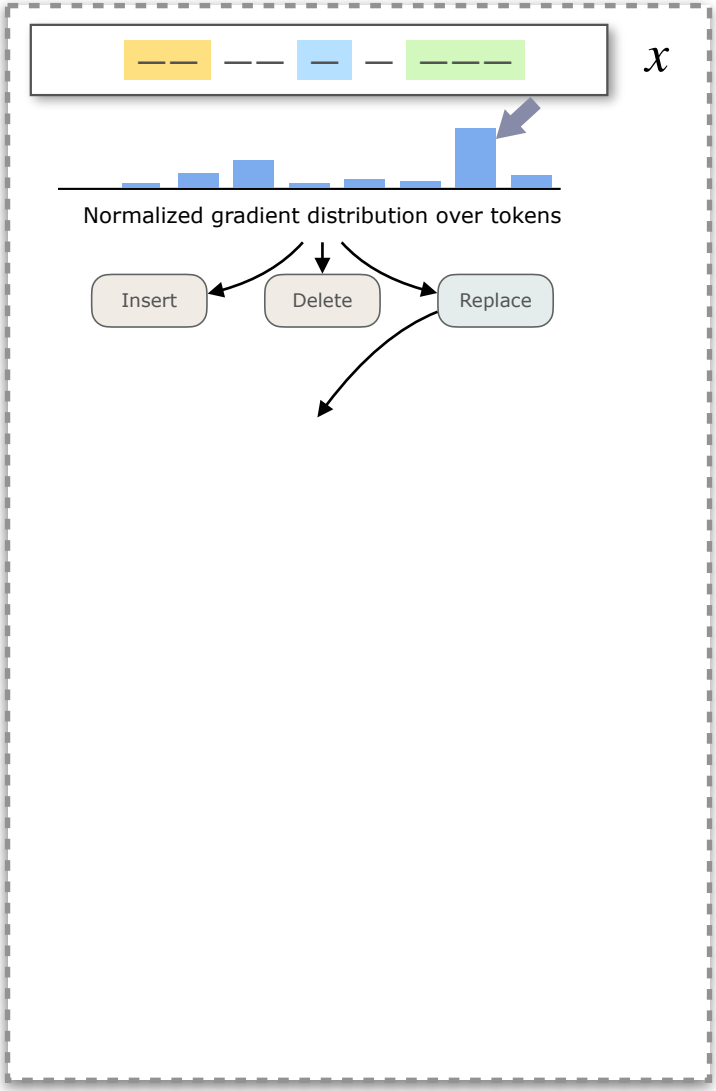
Input: x

↓

Sample a position to edit with a verifier 

↓

Choose an editing action



Workflow

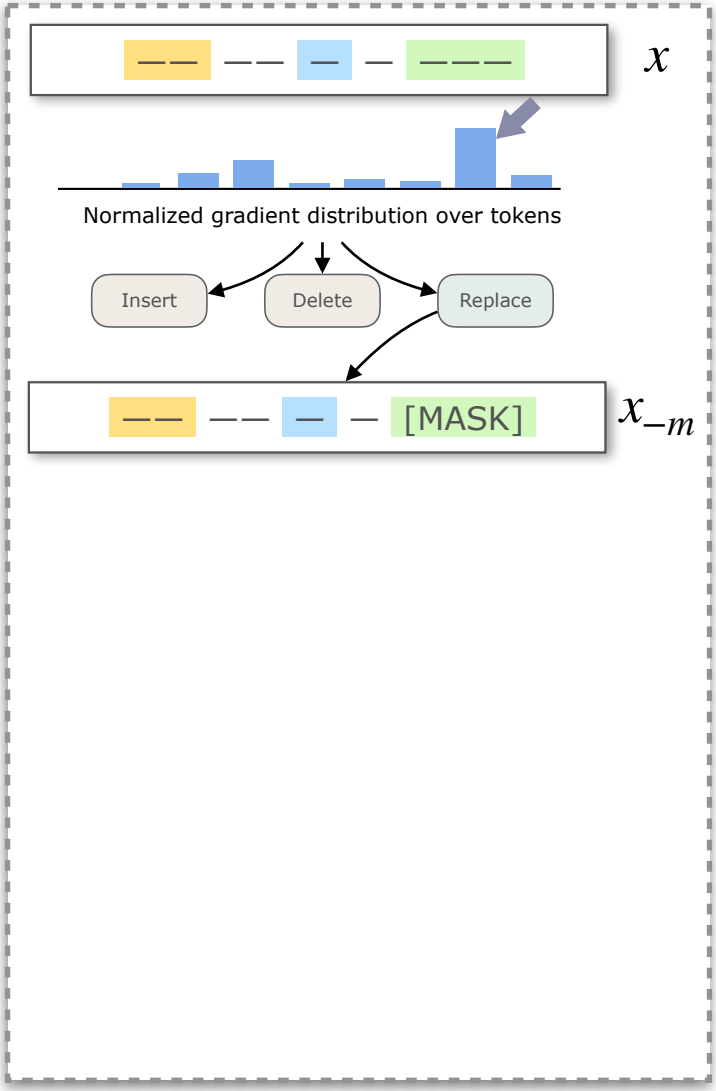
Input: x

↓

Sample a position to edit with a verifier 

↓

Choose an editing action



Workflow

Input: x

↓

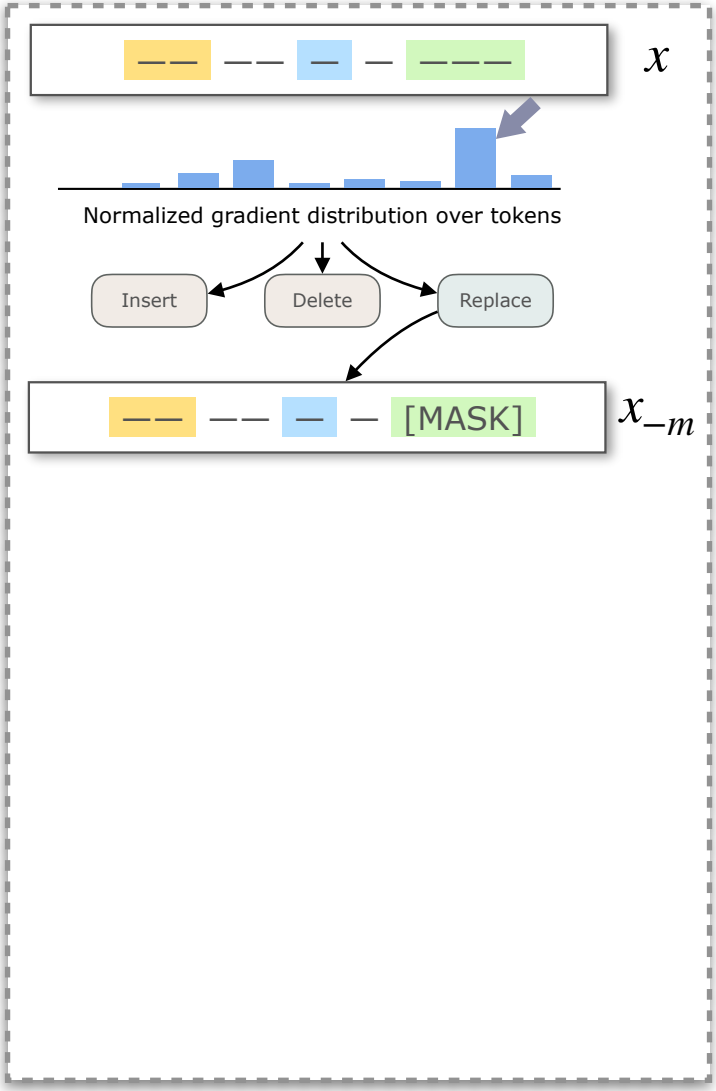
Sample a position to edit with a verifier

↓

Choose an editing action

↓

Sample a token/entity



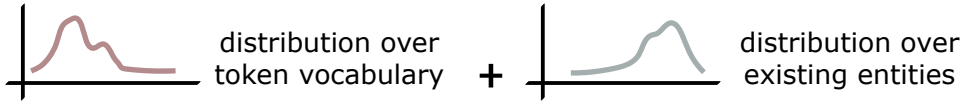
Challenges brought by Multi-token Entity Masking

- The Markov chain must satisfy ***detailed balance condition*** for the sampling to converge on $\pi(x)$.

$$\pi(x')g(x \mid x') = \pi(x)g(x' \mid x)$$

- ***Price for editing entities***: predicting a list of tokens as entities given one mask is not reversible.
  - *Broken entity* will *never* be reached during a multi-token *insertion*, but will be during a token-by-token *deletion*.
  - e.g. *Will Smith starred in The Pursuit for.*

Solutions

- **Separating** the sampling space into a token space and an entity space
  - 

- **Generative proposal** with a T5 model
  - Training of T5: fine-tuned on self-supervised data of token/ entity mask-predicting on Supported claims.
  - Multi-tasking of T5: different prefixes.
- **Reversible**
  - A delete+insert action combination can communicate two spaces.
  - e.g., delete(entity) + insert(token)

Generative Proposal Model $P_3(x' \mid x_{-m}, a)$: Token vs. Entity

- **Replacement**

$$P_3(x' \mid x_{-m}, \texttt{rep}) = \begin{cases} P_3^{\text{ent}}(x' \mid x_{-m}, \texttt{rep}), & x_m = \texttt{ent} \\ P_3^{\text{tok}}(x' \mid x_{-m}, \texttt{rep}), & x_m = \texttt{tok} \end{cases}$$

- **Insertion**

$$P_3(x' \mid x_{-m}, \texttt{ins}) = \alpha P_3^{\text{tok}}(x' \mid x_{-m}, \texttt{ins}) + (1-\alpha) P_3^{\text{ent}}(x' \mid x_{-m}, \texttt{ins})$$

- **Deletion** (reverse of insertion)

$$P_3(x' \mid x_{-m}, \texttt{del}) = 1$$

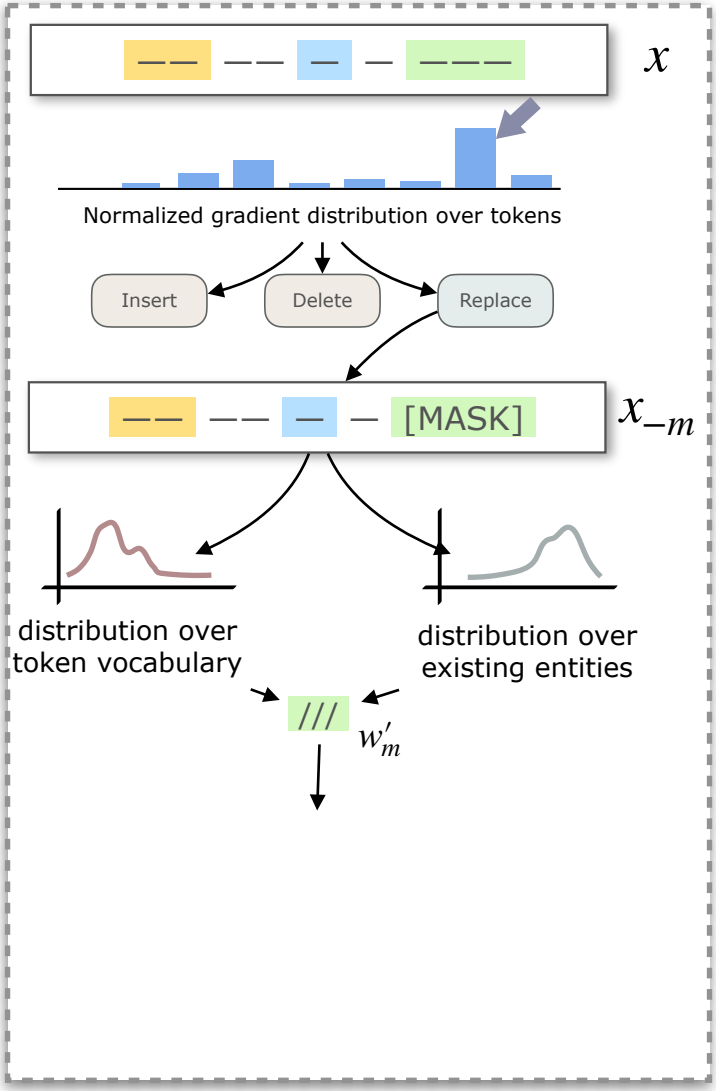
Workflow

Input: x

Sample a position to edit with a verifier

Choose an editing action

Sample a token/entity



Workflow

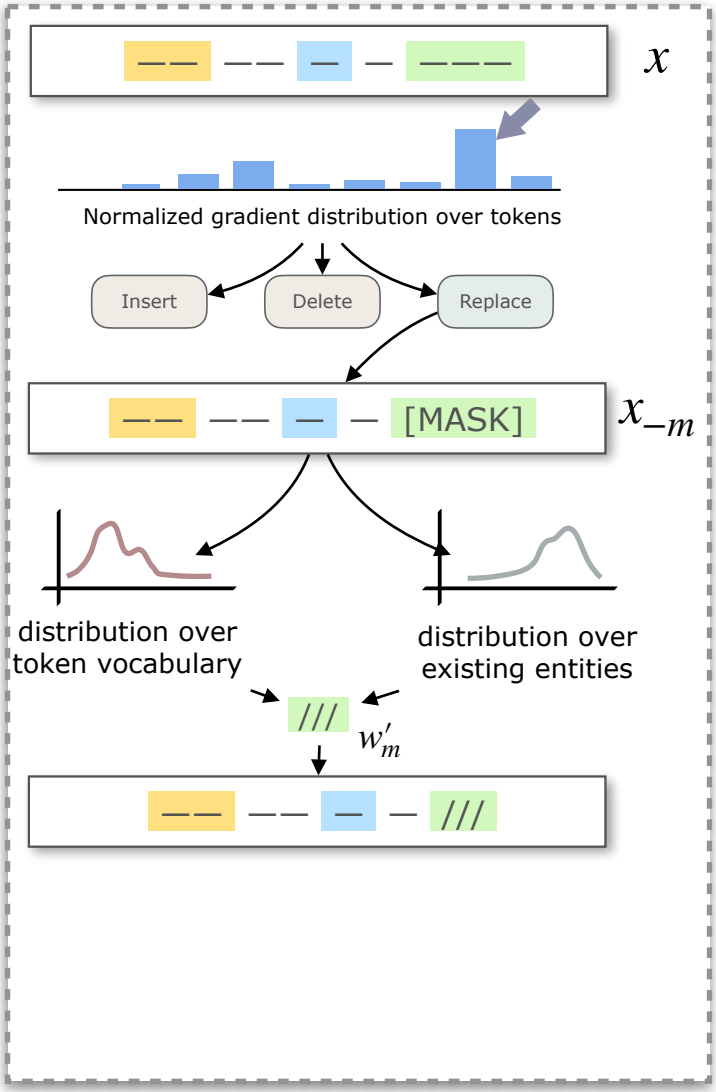
Input: x

Sample a position to edit with a verifier

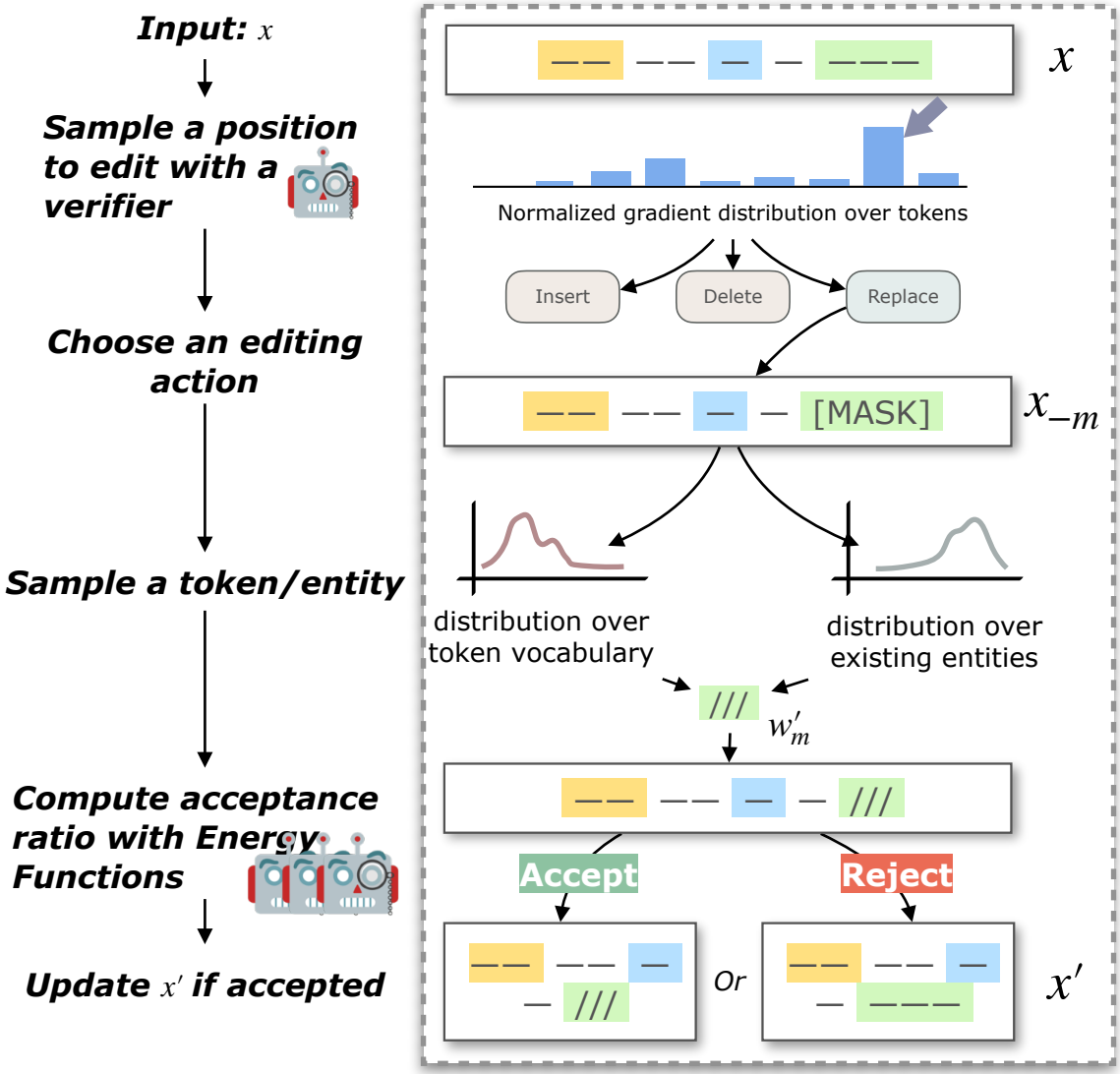
Choose an editing action

Sample a token/entity

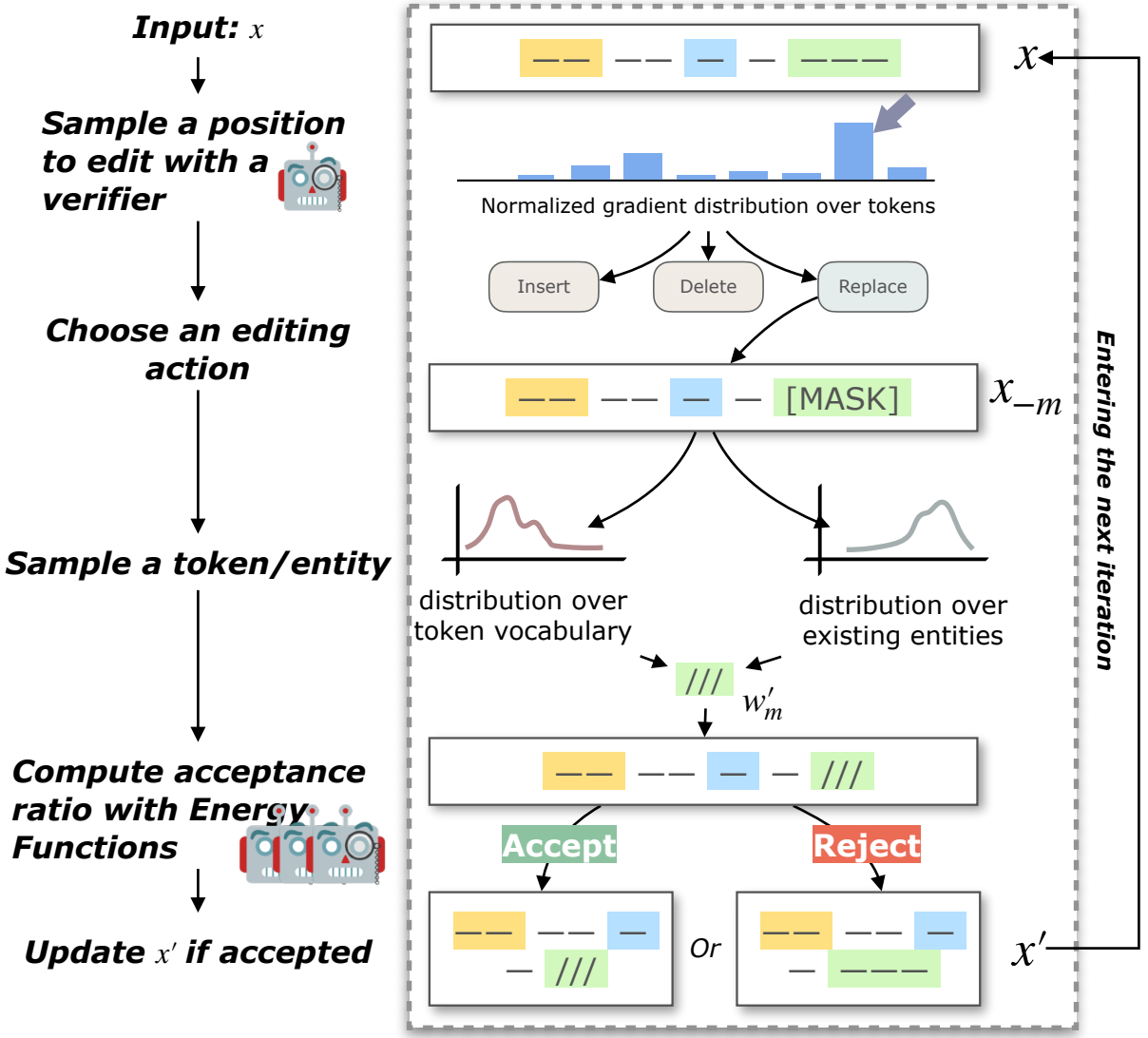
Compute acceptance ratio with Energy Functions



Workflow


Input: x
Sample a position to edit with a verifier
Choose an editing action
Sample a token/entity
Compute acceptance ratio with Energy Functions
Update x' if accepted
x
Normalized gradient distribution over tokens
Insert
Delete
Replace
[MASK]
x_{-m}
distribution over token vocabulary
distribution over existing entities
///
w'_m
Accept
Reject
Or
x'

Workflow


Input: x
Sample a position to edit with a verifier
Choose an editing action
Sample a token/entity
Compute acceptance ratio with Energy Functions
Update x' if accepted
Normalized gradient distribution over tokens
Insert
Delete
Replace
[MASK]
x
x_{-m}
distribution over token vocabulary
distribution over existing entities
///
w'_m
Accept
Reject
Or
x'
Entering the next iteration

Experiments: Setup

- **Datasets**
  - FECData for FEC (Thorne et al. 2021)
  - FEVER for FV (Thorne et al. 2018)

Label	# Train	# Valid	# Test
SUPPORTED	37,961	1,477	1,593
REFUTED	20,075	2,091	2,289

Table 1: Statistics of FECDATA (Thorne and Vlachos 2021), with data sample counts of each split and label.

- **Metrics**
  - SARI scores: the F1 of words being *added*/*deleted*/*kept*
  - Human evaluation (accuracy)

Experiments: Baselines

- **Supervised baselines**
  - T5 (Raffel et al. 2020)
  - EdiT5 (Mallinson et al. 2022)
- ***Distantly-supervised baselines***
  - **DS-1**: Train to propose with evidence-based mask-prediction.
    - *MLM* (Devlin et al. 2019);
    - *2EntPtr* (Shah et al. 2020);
    - ***T5MC*** (Thorne et al. 2021)
  - **DS-2**: Give a verifier (external discriminative models), e.g., NLI, FV.
    - ***T5MC-V*** (Thorne et al. 2021)

Significant Improvements over Previous DS SoTA

Method	Verifier	SARI (%)				RG-2
		Keep	Delete	Add	Final	
Fully Supervised						
T5-base	-	79.6	90.2	59.2	76.4	72.7
EdiT5-base	-	**81.8**	**93.0**	**63.4**	**79.4**	**76.9**
Distantly Supervised						
MLM	-	56.1	52.9	7.8	38.9	42.7
2EncPtr	BERT_b	34.5	48.1	1.7	28.1	34.8
T5MC	-	65.2	62.7	15.5	47.8	50.3
+*enumerate*	BERT_b	66.2	**64.3**	17.1	49.2	51.2
T5MC-V	BERT_b	61.1	54.3	19.4	44.9	42.0
+*enumerate*	BERT_b	63.0	55.7	24.1	47.6	45.5
VENCE	BERT_b	66.0	60.1	34.8	53.6	57.7
	RoBERTa_l	**67.1**	61.9	**36.0**	**55.0**	**59.1**

Table 2: The automatic evaluation results of VENCE compared with baselines. Distantly supervised methods with verifiers apply the DS−2 strategy during training. Verifier_b and Verifier_l denote the base and large version of the pre-trained language model that the Verifier uses.

- VENCE makes better use of the verifier.
- VENCE adds more sensical tokens than baselines.
- Still far behind supervised methods.

Part I: Analysis on Constraints

How does verification affect correction?

Dataset	Verifier	Acc (%)	SARI (%)			
			Keep	Delete	Add	Final
MultiNLI	BERT_b	84.6	63.9	57.0	30.1	50.3
	RoBERTa_l	**90.2**	**65.1**	**58.9**	**32.2**	**52.0**
FEVER	BERT_b	71.7	66.0	60.1	34.8	53.6
	RoBERTa_l	**72.9**	**67.1**	**61.9**	**36.0**	**55.0**

Table 3: Ablation results of the verification model used in VENCE w.r.t. model sizes and trained datasets. We report accuracy of verifier models on test set of MultiNLI and FEVER, respectively.

- Better verifier leads to better performance.
- Even OOD verifier (NLI) helps VENCE outperform baselines.

How do different combinations of energy functions affect correction?

Row	Verif.	Hamm.	LM	SARI-F ↑	HammS. ↓	BERTS. ↑
1	✓			41.2	3.42	0.90
2		✓		25.3	2.08	0.76
3			✓	37.1	**1.19**	0.82
4		✓	✓	39.8	2.10	0.87
5	✓		✓	49.7	2.36	0.92
6	✓	✓		45.0	2.30	0.89
7	✓	✓	✓	**53.6**	2.21	**0.93**

Table 4: Correction results of VENCE with different combinations of energy functions.

- Verified truthfulness score contributes the most 👍 among the three energy scores.

How do different combinations of energy functions affect correction?

Row	Verif.	Hamm.	LM	SARI-F ↑	HammS. ↓	BERTS. ↑
1	✓			41.2	3.42	0.90
2		✓		25.3	2.08	0.76
3			✓	37.1	**1.19**	0.82
4		✓	✓	39.8	2.10	0.87
5	✓		✓	49.7	2.36	0.92
6	✓	✓		45.0	2.30	0.89
7	✓	✓	✓	**53.6**	2.21	**0.93**

Table 4: Correction results of VENCE with different combinations of energy functions.

- When only using $\mathscr{E}_{\mathrm{LM}}$, the claim is not edited much, thus achieving the best HammingScore 👍 but poor SARI-F 👎.

How do different combinations of energy functions affect correction?

Row	Verif.	Hamm.	LM	SARI-F ↑	HammS. ↓	BERTS. ↑
1	✓			41.2	3.42	0.90
2		✓		25.3	2.08	0.76
3			✓	37.1	**1.19**	0.82
4		✓	✓	39.8	2.10	0.87
5	✓		✓	49.7	2.36	0.92
6	✓	✓		45.0	2.30	0.89
7	✓	✓	✓	**53.6**	2.21	**0.93**

Table 4: Correction results of VENCE with different combinations of energy functions.

- SARI-F drops ⬇️ when removing $\mathscr{E}_{\mathrm{LM}}$, showing importance of fluency constraint in correction.

How do different combinations of energy functions affect correction?

Row	Verif.	Hamm.	LM	SARI-F ↑	HammS. ↓	BERTS. ↑
1	✓			41.2	3.42	0.90
2		✓		25.3	2.08	0.76
3			✓	37.1	**1.19**	0.82
4		✓	✓	39.8	2.10	0.87
5	✓		✓	49.7	2.36	0.92
6	✓	✓		45.0	2.30	0.89
7	✓	✓	✓	**53.6**	2.21	**0.93**

Table 4: Correction results of VENCE with different combinations of energy functions.

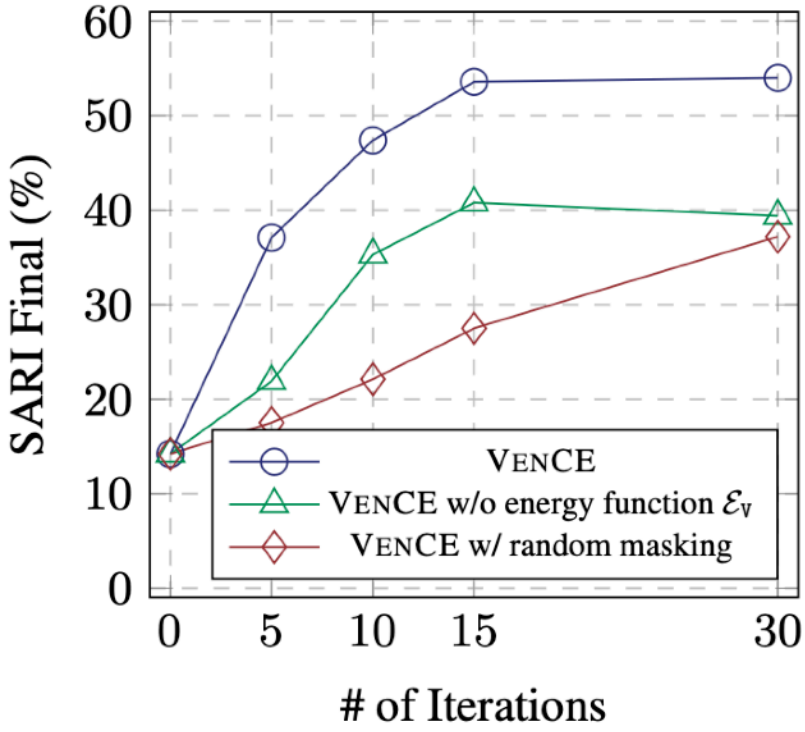
- $\mathscr{E}_{\mathrm{H}}$ does not contributes to the Hamming distance score, but rather improves SARI for correctness.

Part II: Analysis on Editing

Case Study: Editing History

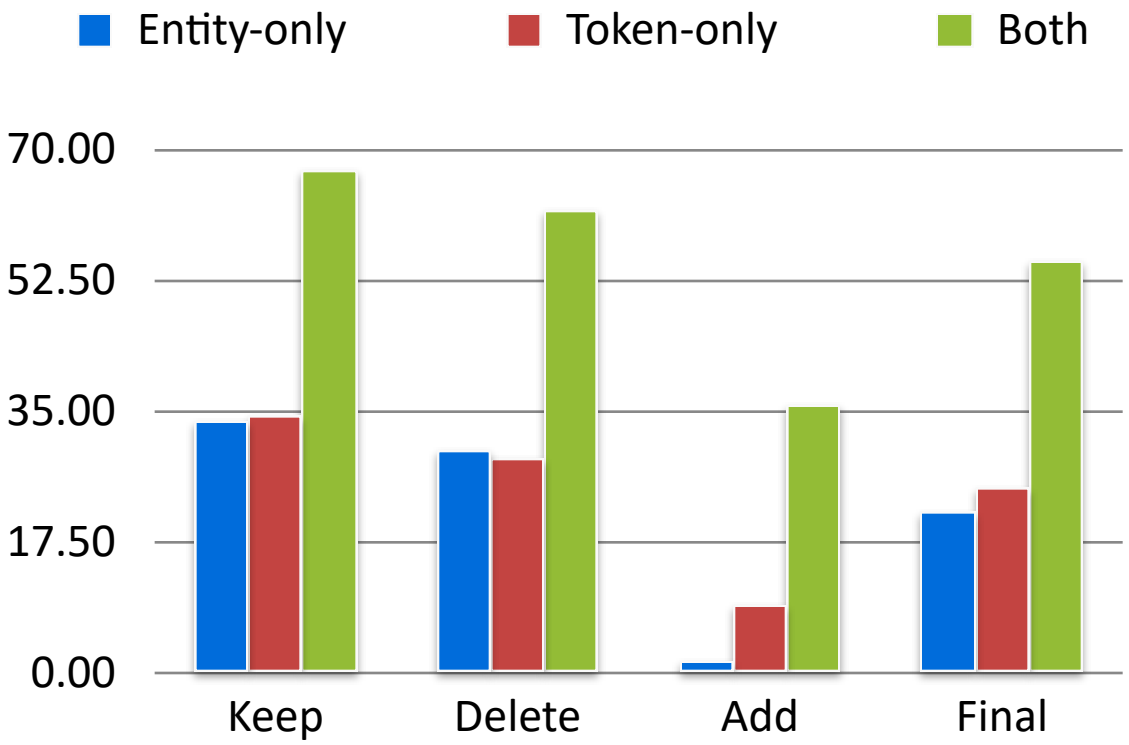
# Iter.	Proposed Claims	Acc.
0	One True Thing is a German film.	-
1	One True Thing is a film.	✓
2	One True Thing is film.	✗
3	One True Thing is a drama film.	✓
4	One True Thing is drama film.	✗
5	One True Thing is American drama film.	✗
6	One True Thing is American drama film.	✓
7	One True Thing is an American drama film.	✓
8	One True Thing is an American film.	✗
9	One True Thing is an American film.	✓
10	One True Thing is an American drama.	✗
11	One True Thing is an 1998 American film.	✗
12	One True Thing is an American.	✗
13	**One True Thing is an American drama film.**	✓
14	One True Thing is an American drama.	✗
15	One True Thing is an American film.	✗
16	One True Thing is a American drama film.	✗
17	One True Thing is an American film.	✗
18	One True Thing is an 1998 American drama film.	✗
19	Anna Quindlen is an American drama film.	✗
20	One True Thing is an American real film.	✗
-	One True Thing is an American film.	Gold

Will more editing iterations help correction?



- VENCE converges at Iter #15.
- Performance drop when losing $\mathscr{E}_V(x)$ (fact verification).
- Gradient-based sampling accelerates convergence.

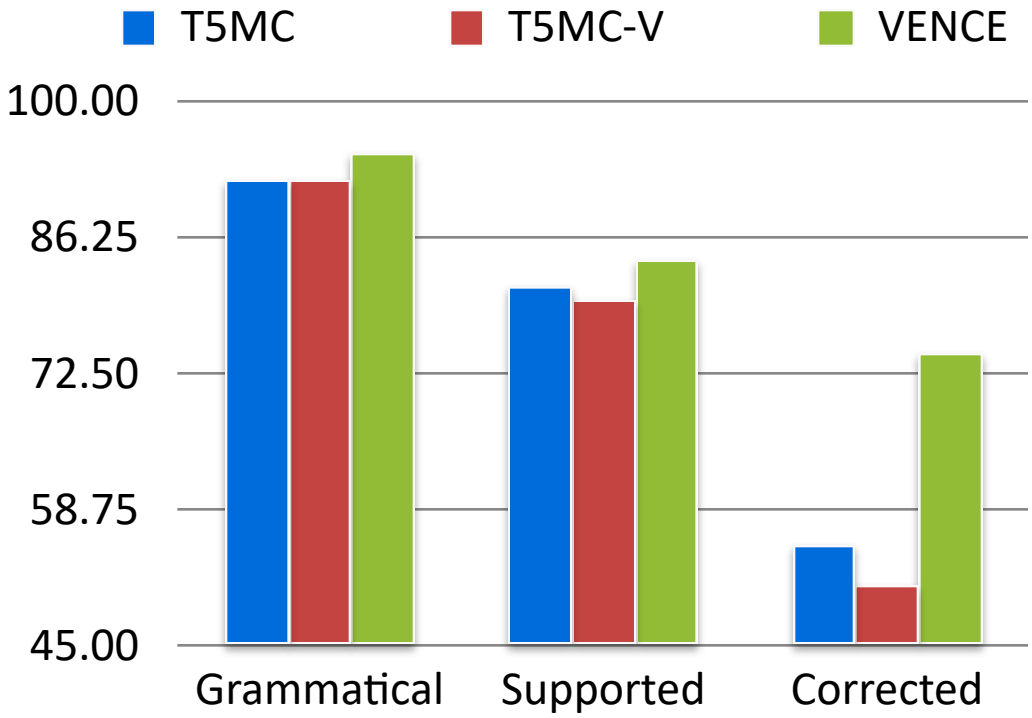
What if we only edit tokens or entities?



- Generative proposal over two spaces greatly surpasses the token-only or entity-only counterparts.

Part III: Manual Evaluation

Manual Evaluation



1. *Is it **grammatically** correct?*
2. *Is it **supported** by evidence?*
3. *Are the original errors **corrected**?*

- Both VENCE and baselines can recover reasonable results from evidence to the output (**supported**).
- VENCE better concentrates on the correction of given errors (**corrected**).

Case Study: *Supported* but over-/not *Corrected*

Claim #1: Shut Up is by hell yeah.

Evidence #1-1: [Shut Up (The Black Eyed Peas song)] "Shut Up" is a song by the American group The Black Eyed Peas. The song was released as the second single from their studio album, Elephunk.
Evidence #1-2: [Shut Up (R. Kelly song)] "Shut Up" is a song written and performed by American R&B singer R. Kelly and included on his twelfth solo studio album Black Panties. The song was released on November 10, 2011 through YouTube,The song was the first song he recorded after his throat."

Correction by T5MC	**Correction by T5MC-V**	**Correction by VENCE**
Shut Up is by the American group The Black Eyed Peas.	Shut Up is by the American group The Black Eyed Peas.	Shut Up is by R. Kelly.
⇒ SUPPORTED BUT OVER-CORRECTED	⇒ SUPPORTED BUT OVER-CORRECTED	⇒ SUPPORTED AND CORRECTED

Claim #2: Aphrodite is unmarried.

Evidence #2-1: [Aphrodite] and had many lovers — both gods, such as Ares, and men, such as Anchises. She played a role in the Eros and Psyche legend, and was both lover and surrogate mother of Adonis. Many lesser beings were said to be children of Aphrodite.
Evidence #2-2: [Aphrodite] claimed to be her place of birth. In Greek myth, the other gods feared that Aphrodite's beauty might lead to conflict and war, through rivalry for her favours; so Zeus married her off to Hephaestus. Despite this, Aphrodite followed her own inclinations.

Correction by T5MC	**Correction by T5MC-V**	**Correction by VENCE**
Aphrodite is both lover and surrogate mother of Adonis.	Aphrodite had many lovers.	Aphrodite is married.
⇒ SUPPORTED BUT NOT CORRECTED	⇒ SUPPORTED BUT NOT CORRECTED	⇒ SUPPORTED AND CORRECTED

Takeaways

- **VENCE** uses **fact verification constrained correction** as a key component in its iterative editing process.
  - Further analysis demonstrates that truthfulness signals from verification serve as an indispensable part of the approach.
- **Finding sibling tasks** and **leveraging techniques developed for those tasks** can be a useful strategy for tackling tasks with limited training data.
- **Limitations**
  - External fact verification models are not able to distinguish between different degrees of factual errors.
  - More comprehensive datasets for the FEC task.

Thanks 😊

Find more about our work via ⬇️

📧 jjchen19@fudan.edu.cn

https://jiangjiechen.github.io/publication/vence/