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

Jiangjie Chen, Rui Xu, Wenxuan Zeng,
Changzhi Sun, Lei Li, Yanghua Xiao
James Webb Space Telescope took the very first pictures of a planet outside of our own solar system.

In 2023, the JWST spotted a number of galaxies nicknamed “green peas.” They were given this name because they are small, round, and green, like peas.

The telescope captured images of galaxies that are over 13 billion years old. This means that the light from these galaxies has been traveling for over 13 billion years to reach us.

JWST took the very first pictures of a planet outside of our own solar system. These distant worlds are called “exoplanets.” Exo means “from outside.”

These discoveries can spark a child’s imagination about the infinite wonders of the universe.
A Factual Error by Bard AI Chatbot Just Cost Google $100 Billion

February 9 2023

https://time.com/6254226/alphabet-google-bard-100-billion-ai-error/
A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity

Yejin Bang
Holy Lovenia
Samuel Cahyawijaya
Nayeon Lee
Wenliang Dai
Dan Su
Bryan Wilie
Ziwei Ji
Tiezhen Yu
Willy Chung
Quyet V. Do
Yan Xu
Pascale Fung

Centre for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
yjbang@connect.ust.hk, pascale@ece.ust.hk
Claim

Will Smith starred in The Truman Show in 2006.

Retrieved Evidence

The Pursuit of Happyness is a 2006 American biographical drama film directed by Gabriele Muccino and starring Will Smith as Chris Gardner, a homeless ...

Corrected Claim

Will Smith starred in The Pursuit for Happyness in 2006.
How to Correct Claims?

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

Will Smith starred in The Truman Show in 2006. ||

Mask spans

Will Smith starred in [MASK] [MASK] [MASK] in 2006. ||

LM generation

Will Smith starred in The Pursuit for Happiness Show in 2006.

No supervised data?

using Fact verification model to rank!

James Thorne, Andreas Vlachos, Evidence-based Factual Error Correction. ACL 2021.
Challenges

• **Semantic Drift of Intended Meaning**
  
  - “New York has 18 million population.” => “New York is in the US.”

  ✴ => *enforcing minimal-edits*

• **Lack Annotated Dataset for FEC**
  
  - High-quality datasets are costly to build.
  - Most datasets are synthetic.

  ✴ => *Self-supervised editing proposal*
Will Smith starred in The Pursuit for Happyness in 2006.

Retrieved Evidence

The Pursuit of Happyness is a 2006 American biographical drama film directed by Gabriele Muccino and starring Will Smith as Chris Gardner, a homeless …

Claim

Will Smith starred in The Truman Show in 2006.

Masked Claim

[--- --- --- [MASK]]

Proposed Correction

[--- --- --- 1998]

Corrected Claim

Will Smith starred in The Pursuit for Happyness in 2006.
Desired Properties

Sampling from Target distribution

\[ \pi(x) = \frac{e^{-\mathcal{E}(x)}}{Z} \]

\[ \mathcal{E}(x) = \mathcal{E}_{\text{LM}}(x) + \mathcal{E}_V(x) + \mathcal{E}_H(x) \]

**Fluency:**
- Language Modeling

**Truthfulness:**
- Fact Verification

**Num-edits:**
- Hamming Distance

\[ \mathcal{E}_{\text{LM}}(x) = - \sum_i \log P_{\text{MLM}}(w_i | x_{-i}) \]

\[ \mathcal{E}_V(x) = - \log P_V(\text{Supported} | x, E) \]

\[ \mathcal{E}_H(x) = \text{HammingDistance}(x, x^0) \]
Input: \( x \)  

Will Smith starred in *The Truman Show* in 2006.
Input: $x$

Randomly select a position to edit

Will Smith starred in The Truman Show in 2006.
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.
Randomly select a position to edit

Choose an editing action

Input: \( x \)

Will Smith starred in The Truman Show in 2006.
**Input:** \( x \)

Sample a position to edit with a verifier

Choose an editing action

---

**Will Smith** starred in **The Truman Show** in **2006**.

\[ \begin{align*}
\text{Insert} & \quad \text{Delete} & \quad \text{Replace} \\
\_\_\_ & \quad \_\_\_ & \quad \_\_ \quad [\text{MASK}] \\
\_\_\_ & \quad \_\_\_ & \quad \_\_ \quad \_\_\_ \\
\_\_\_ & \quad \_\_\_ & \quad \_\_ \quad \_\_\_ \\
\_\_\_ & \quad \_\_\_ & \quad \_\_ \quad \_\_\_ \\
\end{align*} \]
VENCE

**Input:** $x$

Sample a position to edit with a verifier

Choose an editing action

Sample a token/entity

**Example:**

Will Smith starred in The Truman Show in 2006.

---

| Insert | Delete | Replace |
|--------|--------|---------|
|        |        |         |

$2006$

$[\text{MASK}]$
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.

---

... The Pursuit for Happyness ...

... The Pursuit for ...

... [MASK] ...

... The Pursuit for ...

---
Solution

- **Separating** the sampling space into a token space and an entity space
  - \[ \text{distribution over token vocabulary} + \text{distribution over existing entities} \]

- **Generative proposal** with a T5 model

- **Reversible**
  - A delete+insert action combination can communicate two spaces.
  - e.g., delete(entity) + insert(token)
Generative Proposal Model $P_3(x'|x_{-m}, a)$: Token vs. Entity

- **Replacement**

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

- **Insertion**

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

- **Deletion** (reverse of insertion)

  $$P_3(x'|x_{-m}, \text{del}) = 1$$
**Sample a position to edit with a verifier**

**Choose an editing action**

**Sample a token/entity**

*Input: $x$*

**Will Smith** starred in **The Truman Show** in **2006**.

---

**Choose an editing action**

**Sample a token/entity**

*Input: $x$*

**Will Smith** starred in **The Truman Show** in **2006**.

---

**Choose an editing action**

**Sample a token/entity**

*Input: $x$*
**Input:** $x$

Sample a position to edit with a verifier

Choose an editing action

Sample a token/entity

---

**Will Smith** starred in **The Truman Show** in **2006**.

---

Choose an editing action:

- Insert
- Delete
- Replace

Sample a token/entity:

- Distribution over token vocabulary
- Distribution over existing entities

---

Input: $x$

Will Smith starred in The Truman Show in 2006.
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

**Will Smith** starred in **The Truman Show** in **2006**.

*Sample a token/entity*
**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

---

**Example:**

Will Smith starred in The Truman Show in 2006.

---

- **Insert**
- **Delete**
- **Replace**

distribution over token vocabulary

distribution over existing entities

Accept

Reject

$w'_m$
Merits

1. No need for supervision from FEC.
2. Verified, minimally edited and fluent.
3. Replaceable components.
# 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 an American drama film.                | ✓    |
| 7       | One True Thing is an American film.                      | ✓    |
| 8       | One True Thing is an American drama.                     | ✗    |
| 9       | One True Thing is an 1998 American film.                 | ✗    |
| 10      | One True Thing is an American.                           | ✗    |
| 11      | **One True Thing is an American drama film.**            | ✓    |
| 12      | One True Thing is an American drama.                     | ✗    |
| 13      | One True Thing is an American film.                      | ✗    |
| 14      | One True Thing is a American drama film.                  | ✗    |
| 15      | One True Thing is an American film.                      | ✗    |
|         |                                                           |      |
|         | One True Thing is an American film.                      | Gold |

**Gold** indicates the final accepted claim.
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)
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) (+enumerate)
- DS-2: Give a verifier (external discriminative models), e.g., NLI, FV.
  - T5MC-V (Thorne et al. 2021) (+enumerate)
Significant Improvements over Previous DS SoTA

| Method                          | SARI-Final | ROUGE-2 |
|---------------------------------|------------|---------|
| MLM (Devlin et al. 2019)        | 37.50      | 48.75   |
| 2EncPtr (Shah et al. 2020)      | 48.75      | 60.00   |
| T5MC (Thorne and Vlachos 2021) | 52.50      | 65.00   |
| T5MC + enumerate                | 57.50      | 67.50   |
| T5MC-V (Thorne and Vlachos 2021)| 55.00      | 65.00   |
| T5MC-V + enumerate              | 60.00      | 70.00   |
| VENCE (RoBERTa large)           | 57.50      | 67.50   |

Legend:
- MLM (Devlin et al. 2019)
- 2EncPtr (Shah et al. 2020)
- T5MC (Thorne and Vlachos 2021)
- T5MC + enumerate
- T5MC-V (Thorne and Vlachos 2021)
- T5MC-V + enumerate
- VENCE (RoBERTa large)
Does better verification leads to better correction?

- ID Results on FEVER
- OOD Results on MultiNLI
Which score contributes the most?

- Leave-one-out
- Only-keep-one
Will more editing iterations help correction?

- VENCE converges at Iter #15.
- Performance drop when losing $\mathcal{E}_V(x)$ (fact verification).
- Gradient-based sampling accelerates convergence.
What if we only edit tokens or entities?
Human Evaluation: VENCE productions are not only supported, but with errors corrected.

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

**Evidence 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:** [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.

Aphrodite is both lover and surrogate mother of Adonis. Aphrodite had many lovers. Aphrodite is married.

**T5MC**

True but semantics change.

**T5MC-V**



**VENCE**

Corrected
Summary and Takeaways

• Formulate the factual error correction as a sampling problem
• VENCE: Iterative editing (MCMC) with self-supervised proposals
• Using external fact verification model (LoREN AAAI 2022) as the guidance
• Limitations and future work
  - different degrees of factual errors.
  - Larger comprehensive datasets for the FEC task.
Converge to the Truth: Factual Error Correction via Iterative Constrained Editing

Jiangjie Chen, Rui Xu, Wenxuan Zeng, Changzhi Sun, Lei Li, Yanghua Xiao

For automatic fact verification: LOREN: Chen et al. AAAI 2022

jjchen19@fudan.edu.cn

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