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]

[1]Fudan University [2]University of Electronic Science and Technology of China

[3]ByteDance AI Lab [4]University of California, Santa Barbara

Project Homepage

Introduction

Task: Factual Error Correction

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.

Factual Errors

Previous Work

- ***Methods***: Evidence-based Factual Error Correction
  - one-pass mask-then-correct generation
- ***Limitations***:
  - Lack fine-grained annotations and high-quality datasets, which are costly to build.
  - Most datasets are synthetically built.

The Motivations of This Work

- 🏆 ***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.

Contributions

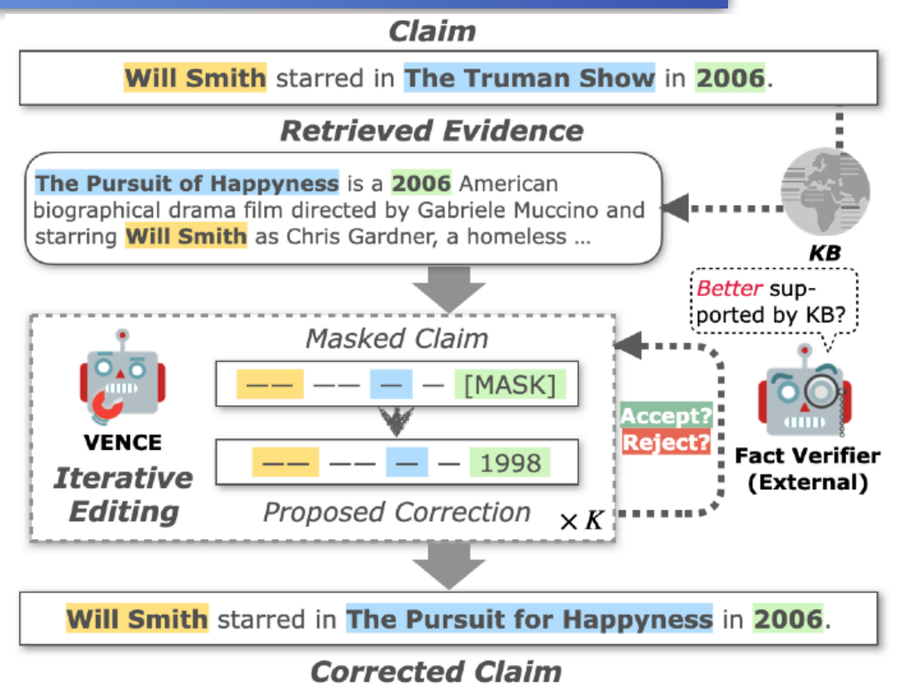
- We are the first to adopt an iterative text editing method (VENCE) for solving factual error correction without direct supervision, which alleviates the over-erasure problem in previous methods
- VENCE enjoys a more powerful error revision ability by effectively integrating external but coarse-grained verification signals during each editing iteration.

The VENCE Framework

Overview of VENCE

No supervised FEC data?

Fact Verification Helps! 🌟



Desired Properties in FEC: Energy Functions

Desired properties of the target texts

$$\mathcal{E}(x) = \mathcal{E}_{\text{LM}}(x) + \mathcal{E}_{\text{V}}(x) + \mathcal{E}_{\text{H}}(x)$$

Fluency
- Language Modeling $\mathcal{E}_{\text{LM}}(x) = -\sum_i \log P_{\text{MLM}}(w_i \mid x_{-i})$

Truthfulness
- Fact Verification $\mathcal{E}_{\text{V}}(x) = -\log P_{\text{V}}(\text{Supported} \mid x, E)$

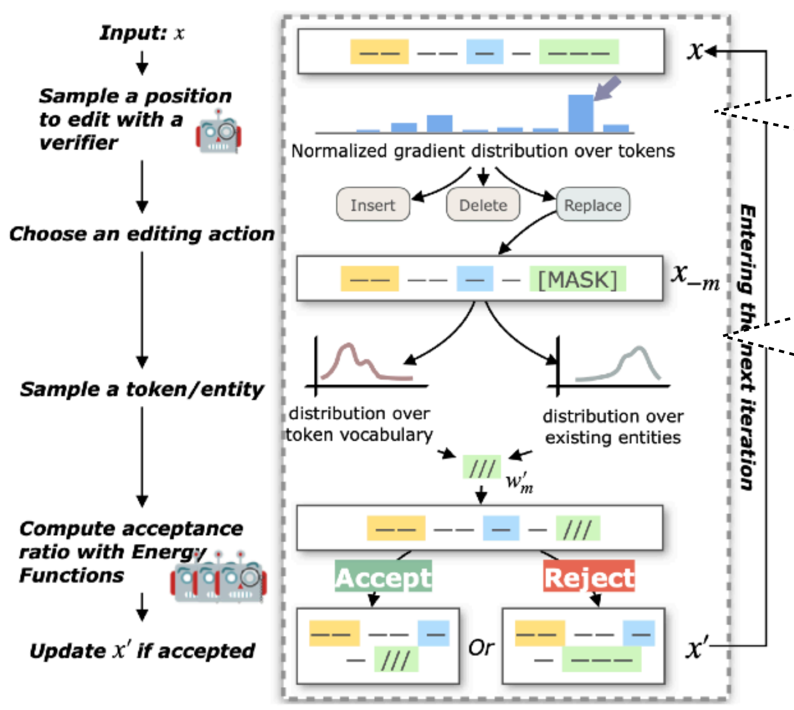
Minimal-edits
- Hamming Distance $\mathcal{E}_{\text{H}}(x) = \text{HammingDistance}(x, x^0)$

Constrained Text Editing via Metropolis-Hastings Sampling

- **Stationary distribution**
  Where we want the sampling to converge $\quad \pi(x) = \dfrac{e^{-\mathcal{E}(x)}}{Z}$
- **Transition distribution** $\quad g(x' \mid x) = P_1(m \mid x)P_2(a)P_3(x' \mid x_{-m}, a)$
  In the Markov chain, taking the action a to edit position m
- **Acceptance Ratio**
  Decides the acceptance of each proposal

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

The workflow of VENCE



Editing Position Proposal
- *Sample a position based on Verifier (P_V)'s normalized gradient distribution.*
- *Multi-token Entity Masking*

Generative Proposal Model (Token vs. Entity)
- *Replacement*

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

- *Insertion*

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

- *Deletion (reverse of insertion)*

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

Experiments

Datasets and Metrics

- **Datasets:** FECData for FEC & FEVER for FV
- **Metrics:** SARI scores & Human evaluation (accuracy)

SARI scores evaluate the F1 of words being added/deleted/kept

Baselines

- **Supervised Baselines:** T5 & EdiT5
- **Distantly-Supervised Baselines:**
  – **DS-1**: Train to propose with evidence-based mask-prediction —> *MLM 2EntPtr T5MC*
  – **DS-2**: Give a verifier (external discriminative models), e.g., NLI, FV. —> *T5MC-V*

The automatic evaluation results of VENCE compared with baselines

Method	Verifier	SARI (%) Keep	Delete	Add	Final	RG-2
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**

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

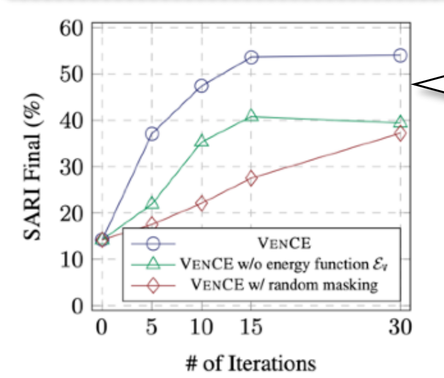
Analysis on Constraints - How do verification affect correction?

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**

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

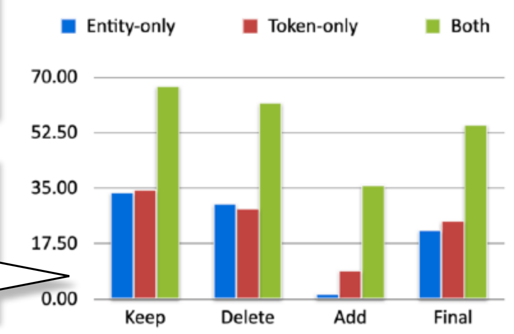
Analysis on Editing

Q1: Will more iterations help correction?



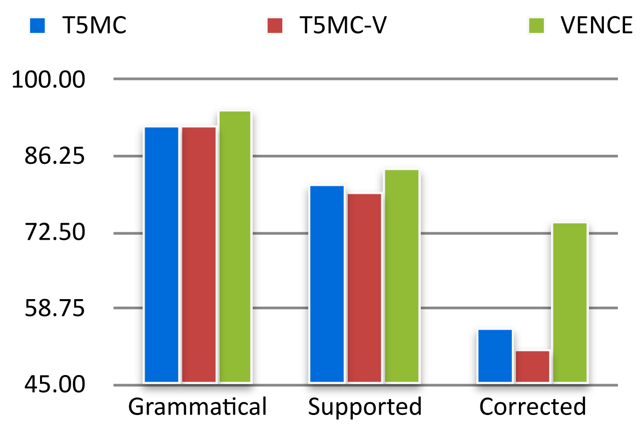
- VENCE converges at Iter #15.
- Performance drop when losing $\mathcal{E}_V(x)$ (fact verification).
- Gradient-based sampling accelerates convergence.

Q2: What if we only edit tokens or entities?



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

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**).

If you have any questions, please email: jjchen19@fudan.edu.cn