DiffusER

Diffusion via **E**dit-based **R**econstruction

Machel Reid[1], Vincent Josua Hellendoorn[2], Graham Neubig[3]

[1]Google Research; [2]Software and Societal Systems Department Carnegie Mellon University;
[3]Language Technologies Institute, Carnegie Mellon University Inspired Cognition;

ICLR 2023

Introduction

Revision and editing are central to how humans produce content;

Humans write and revise emails and papers, gradually produce works of art, and iterate on plans for a project.

Despite this, the most dominant paradigm in text generation is purely autoregressive, producing text left-to-right in a single pass.

- Although models employing this single-pass form of generation are highly performant, they are **Limited by the Inability to Refine Existing Text.**

Introduction (cont.)

Revision and editing are central to how humans produce content;

To address this, this study propose DiffusER: Diffusion via Edit-based Reconstruction, a flexible method to apply edit-based generative processes to arbitrary text generation tasks.

DiffusER is not only a strong generative model in general, rivalling autoregressive models on several tasks spanning machine translation, summarization, and style transfer.

- Can also perform other varieties of generation that standard autoregressive models are not well-suited for.

  e.g. condition generation on a prototype, or an incomplete sequence, and continue revising based on previous edit steps.

DiffusER

Edit-based Generation

The **Corruption and Reconstruction** of DiffusER are based on four **Edit Operations** of {Insert, Delete, Keep, Replace}.

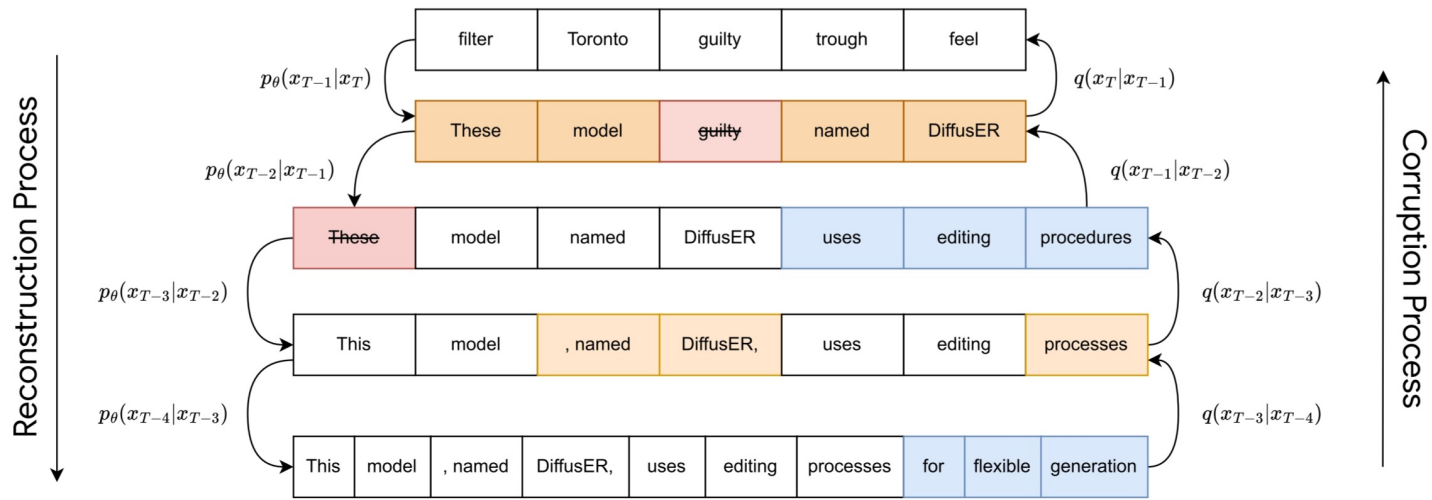


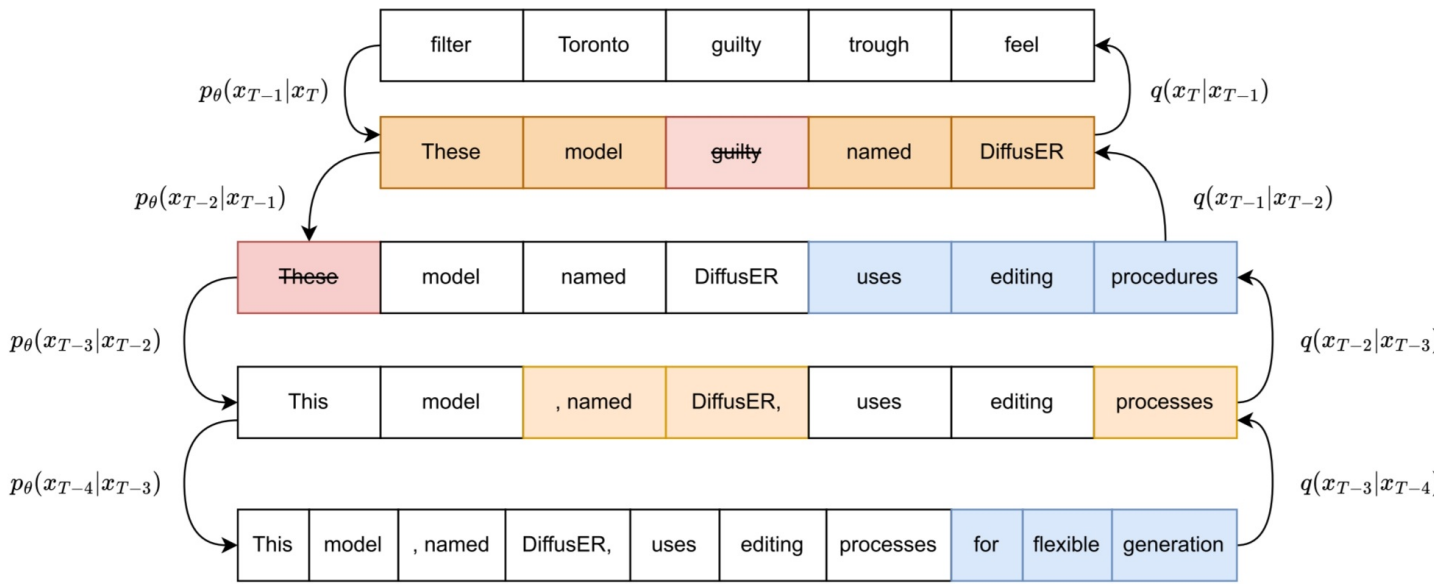
Figure 1: DIFFUSER's text generation process. Orange represents replacements, blue represents insertions, red represents deletions, and white represents keep operations. This process largely imitates a natural editing process (Reid & Neubig, 2022).

DiffusER (cont.)

Edit-based Corruption

Corruption process $q(x_t|x_{t-1}; \mathcal{E}^{tag}, \mathcal{E}^{len})$ is parameterized by two distributions:

- the distribution over edit types $\mathcal{E}^{tag}$ (default 60% keep, 20% replace, 10% delete & 10% insert),
- and the distribution over edit length $\mathcal{E}^{len}$.



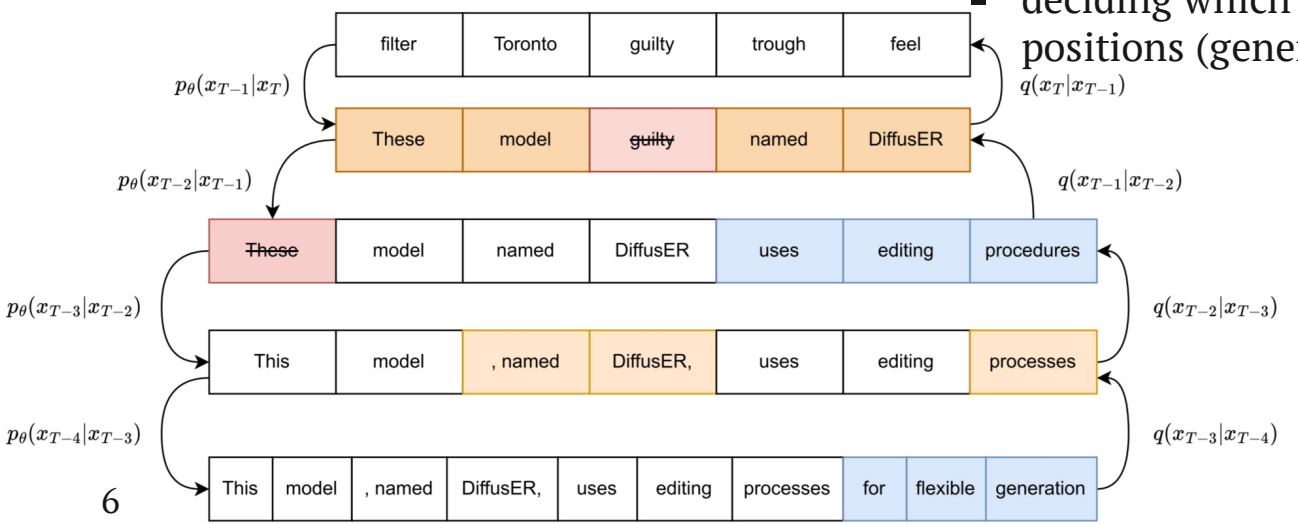
DiffusER (cont.)

Edit-based Reconstruction

$$p_\theta(x_{t-1}|x_t) \sim p_\theta(x_{t-1}, e_t|x_t)$$
$$= p_\theta^{gen}(x_{t-1}|x_t, e_t)p_\theta^{tag}(e_t|x_t)$$

One can think of ER as two distinct steps:

- identify which edits should take place (tagging process) and
- deciding which tokens should go in these positions (generative process).



DiffusER (cont.)

Edit-based Reconstruction

$$p_\theta(x_{t-1}|x_t) \sim p_\theta(x_{t-1}, e_t|x_t) = p_\theta^{gen}(x_{t-1}|x_t, e_t)\underbrace{p_\theta^{tag}(e_t|x_t)}_{\text{predict edit op}}$$

The	dog	is	an	animal	and	is	descended	from	the	wolf
INSERT	**KEEP**	**KEEP**	**REPL**	**DEL**	**DEL**	**DEL**	**REPL**	**REPL**	**KEEP**	**KEEP**

Encoder

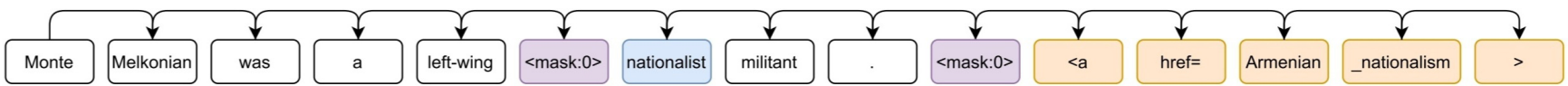
The	dog	is	an	animal	and	is	descended	from	the	wolf

DiffusER (cont.)

Edit-based Reconstruction

$$p_\theta(x_{t-1}|x_t) \sim p_\theta(x_{t-1}, e_t|x_t) = \underbrace{p_\theta^{gen}(x_{t-1}|x_t, e_t)}_{\text{generate Ins. \& Rep.}} p_\theta^{tag}(e_t|x_t)$$

$$<\text{insert sentence:n}> \leftarrow p_\theta^{gen}(\cdots <\text{insert:n}> \cdots <\text{insert:n+1}> \cdots \underbrace{<\text{insert:n}>}_{\textit{predict insertion}})$$



In the generation step, **after Removing tokens selected for Deletion**, they sum a learned embedding to **Insert and Replace types** and generate the inserted and replaced sequences **Autoregressively**.

DiffusER (cont.)

Initialization Techniques & Decoding

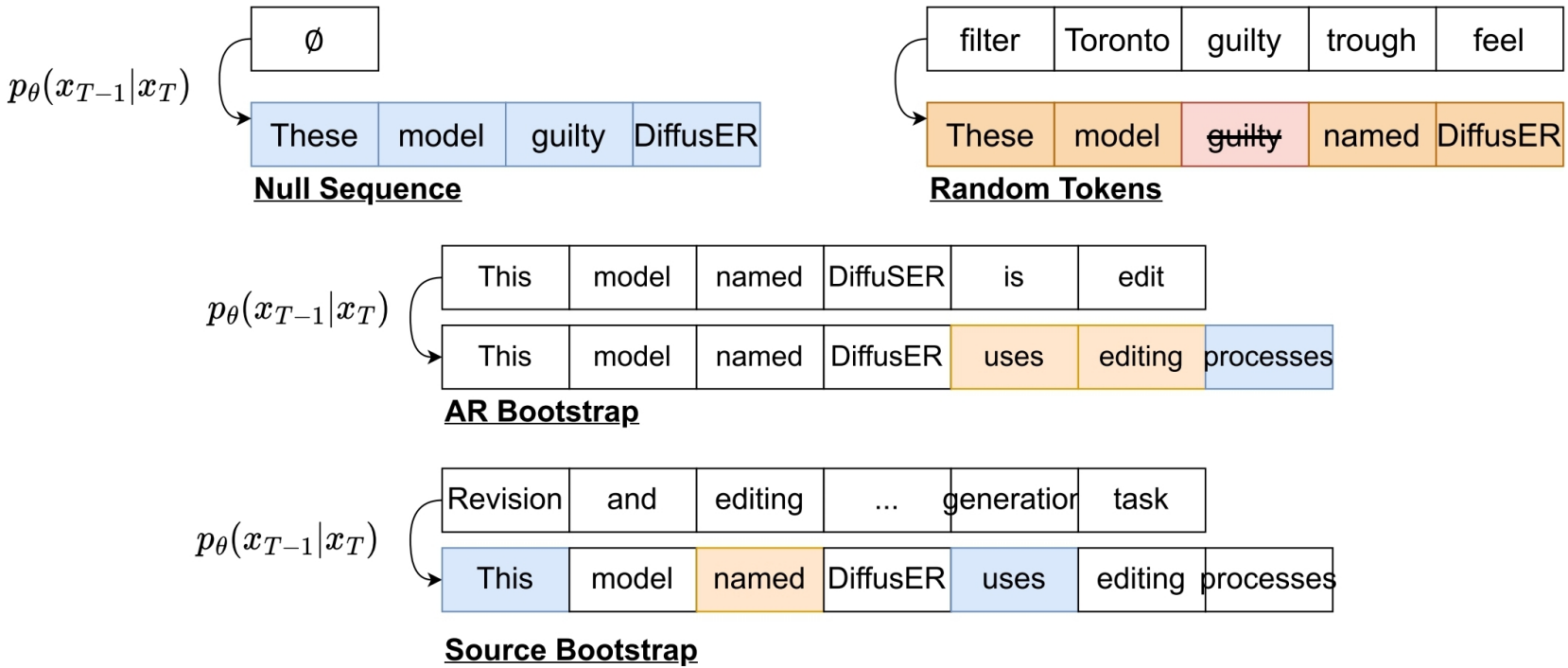


Figure 2: Figure illustrating bootstrapping methods for decoding.

DiffusER (cont.)

Initialization Techniques & **Decoding**

- Greedy
- Beam
- Nucleus
- 2D Beam

  Sequence/Token Level beam width of b & **Revision Level** beam width of r

  - 1. r candidates are fed to the next step of the diffusion model.
    2. For each of r hypotheses the next diffusion step is decoded with beam width of b. (This leads us to have $r \times b$ candidate hypotheses)
    3. Take the top r. This process repeats for each diffusion step thereafter.
  - Increases the beam count to $r \times b$ beams.

Experiments

Machine Translation & Summarization

- Machine Translation

  Use a Poisson distribution $\mathcal{E}^{len}(\lambda = 3)$ over edit operation lengths in corruption process.

- Summarization

  Different in nature from MT, summarization can be described as more conducive to edits as a good summary tends to preserve many parts of the input.

  Use a Poisson distribution $\mathcal{E}^{len}(\lambda = 8)$. **(to roughly model sentence boundaries)**

Experiments (cont.)

Machine Translation & Summarization

Model	En-De (MT)	CNN-DM (Summ)
AR Transformer (Vaswani et al., 2017)	27.3	36.8
SUNDAE (Savinov et al., 2022)	26.3	37.0
CMLM (Ghazvininejad et al., 2019)	24.6	—
Levenshtein Transformer[2] (Gu et al., 2019)	23.7	—
DisCo (**?**)	24.7	—
Imputer	25.2	—
DIFFUSER	27.2	37.8
DIFFUSER + AR bootstrap	**28.8**	38.4
DIFFUSER + source bootstrap	24.5	**38.9**

Table 1: Machine Translation (MT) and Summarization (Summ) results on WMT'14 En-De (gold) and CNN-DailyMail. Experiments on MT use BLEU while summarization uses ROUGE. DIFFUSER is compatible with a standard autoregressive model, while outperforming previous methods.

Experiments (cont.)

Text Style Transfer

Train two separate, style-specific (e.g. positive and negative) DiffusERs on the style-specific data. At test time, e.g.

- positive text → negative DIFFUSER model,
- negative text → positive DIFFUSER model.

Model	Accuracy	BLEU
Masker (Malmi et al., 2020)	40.9	14.5
Tag and Generate (Madaan et al., 2020)	86.2	19.8
LEWIS (Reid & Zhong, 2021)	**93.1**	24.0
DIFFUSER	87.6	**25.2**

Table 2: Results on Yelp dataset for text style transfer. Without task-specific training techniques, DIFFUSER performs comparably to previous task-specific methods.

Experiments (cont.)

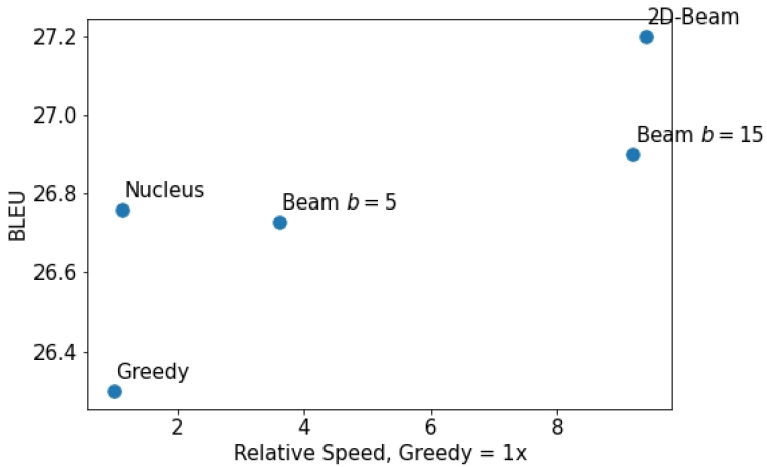


Figure 3: Relative time (seconds) comparison between decoding methods, measured on a single V100 GPU. There is a trade-off between inference cost and performance. Faster well-performing decoding algorithms for diffusion models are an area for further work.

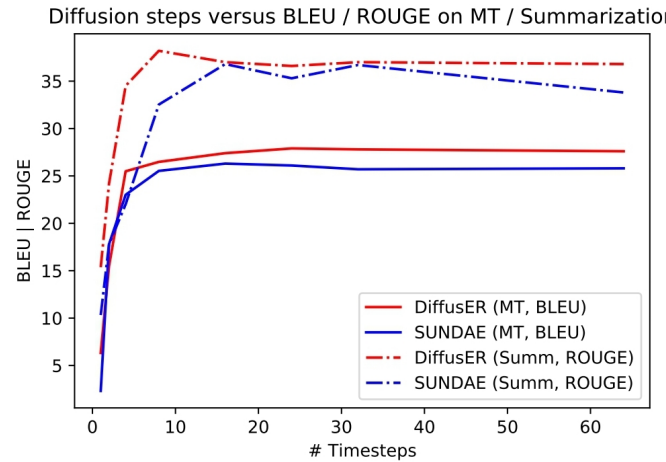


Figure 4: Number of steps versus BLEU/ROUGE on WMT'14 En-De and Summarization for both SUNDAE and DIFFUSER. We observe fast initial progression with performance, leveling off as steps increase.

Conclusions

- Proposed DiffusER, an diffusion-based generative model for text using edits.
- Shows improvements across the tasks considered,

  with improved generative flexibility via incremental text improvement, and compatibility with standard autoregressive models.
- Even without task-specific techniques, DiffusER still has competitive performance with state of the art methods.