Text Editing as Imitation Game

* Ning Shi, Bin Tang, Bo Yuan, Longtao Huang
Yewen Pu, Jie Fu, † Zhouhan Lin

Alberta Machine Intelligence Institute, Dept. of Computing Science, University of Alberta
Alibaba Group
Shanghai Jiao Tong University
Autodesk Research
Beijing Academy of Artificial Intelligence
ning.shi@ualberta.ca, {tangbin.tang, qiufu.yb, kaiyang.hlt}@alibaba-inc.com
yewen.pu@autodesk.com, fujie@baai.ac.cn, lin.zhouhan@gmail.com

Abstract

Text editing, such as grammatical error correction, arises naturally from imperfect textual data. Recent works frame text editing as a multi-round sequence tagging task, where operations – such as insertion and substitution – are represented as a sequence of tags. While achieving good results, this encoding is limited in flexibility as all actions are bound to token-level tags. In this work, we reformulate text editing as an imitation game using behavioral cloning. Specifically, we convert conventional sequence-to-sequence data into state-to-action demonstrations, where the action space can be as flexible as needed. Instead of generating the actions one at a time, we introduce a dual decoders structure to parallel the decoding while retaining the dependencies between action tokens, coupled with trajectory augmentation to alleviate the distribution shift that imitation learning often suffers. In experiments on a suite of Arithmetic Equation benchmarks, our model consistently outperforms the autoregressive baselines in terms of performance, efficiency, and robustness. We hope our findings will shed light on future studies in reinforcement learning applying sequence-level action generation to natural language processing.

1 Introduction

Text editing (Malmi et al., 2022) is an important domain of processing tasks to edit the text in a localized fashion, applying to text simplification (Agrawal et al., 2021), grammatical error correction (Li et al., 2022), punctuation restoration (Shi et al., 2021), to name a few. Neural sequence-to-sequence (seq2seq) framework (Sutskever et al., 2014) establishes itself as the primary approach to text editing tasks, by framing the problem as machine translation (Wu et al., 2016). Applying a seq2seq modeling has the advantage of simplic-

* Work was done at Alibaba Group.
† Zhouhan Lin is the corresponding author.
such as deletion or insertion after a token (Awasthi et al., 2019; Malmi et al., 2019).

In contrast, alternative approaches (Gupta et al., 2019) train the agent to explicitly generate free-
form edit actions and iteratively reconstructs the text during the interaction with an environment
able of altering the text based on these actions. This sequence-level action generation (Branavan
et al., 2009; Guu et al., 2017; Elgohary et al., 2021) allows higher flexibility of action design not limited
to token-level actions, and is more advantageous given the narrowed problem space and dynamic
context in the edit (Shi et al., 2020).

The mechanisms of sequence tagging and sequence generation against end-to-end are exemplified in Figure 1. Both methods allow multiple rounds of sequence refinement (Ge et al., 2018; Liu
et al., 2021) and imitation learning (IL) (Pomerleau, 1991). Essentially an agent learns from the
demonstrations of an expert policy and later imitates the memorized behavior to act independently
(Schaal, 1996). On the one hand, IL in sequence tagging functions as a standard supervised learning
in its nature and thus has attracted significant interest and been widely used recently (Agrawal et al.,
2021; Yao et al., 2021; Agrawal and Carpuat, 2022), achieving good results in the token-level action
generation setting (Gu et al., 2019; Reid and Zhong, 2021). On the other hand, IL in sequence-level
action generation is less well defined even though its principle has been followed in text editing (Shi
et al., 2020) and many others (Chen et al., 2021). As a major obstacle, the training is on state-action
demonstrations, where the encoding of the states and actions can be very different (Gu et al., 2018).
For instance, the mismatch of the lengths dimension between the state and action makes it tricky to implement for an auto-regressive modeling that benefits from a single, uniform representation.

To tackle the issues above, we reformulate text editing as an imitation game controlled by a
Markov Decision Process (MDP). To begin with, we define the input sequence as the initial state, the
required operations as action sequences, and the output target sequence as the goal state. A learning
agent needs to imitate an expert policy, respond to seen states with actions, and interact with the environ-
ment until the success of the eventual editing. To convert existing input-output data into state-
action pairs, we utilize trajectory generation (TG), a skill to leverage dynamic programming (DP) for
an efficient search of the minimum operations given a predefined edit metric. We backtrace explored
editing paths and automatically express operations as action sequences. Regarding the length misalign-
ment, we first take advantage of the flexibility at the sequence-level to fix actions to be of the same
length. Secondly, we employ a linear layer after the encoder to transform the length dimension of the
context matrix into the action length. By that, we introduce a dual decoders (D2) structure that not
only parallels the decoding but also retains capturing interdependencies among action tokens. Taking
a further step, we propose trajectory augmentation (TA) as a solution to the distribution shift problem
most IL suffers (Ross et al., 2011). Through a suite of three Arithmetic Equation (AE) benchmarks
(Shi et al., 2020), namely Arithmetic Operators Restoration (AOR), Arithmetic Equation Simplifi-
cation (AES), and Arithmetic Equation Correction (AEC), we confirm the superiority of our learning
paradigm. In particular, D2 consistently exceeds standard autoregressive models from performance,
efficiency, and robustness perspectives.

In theory, our methods also apply to other imitation learning scenarios where a reward function
exists to further promote the agent. In this work, we primarily focus on a proof-of-concept of our
learning paradigm landing at supervised behavior cloning (BC) in the context of text editing. To this
end, our contributions\textsuperscript{1} are as follows:

1. We frame text editing into an imitation game formally defined as an MDP, allowing the highest
degrees of flexibility to design actions at the sequence-level.
2. We involve TG to translate input-output data to state-action demonstrations for IL.
3. We introduce D2, a novel non-autoregressive decoder, boosting the learning in terms of accuracy, efficiency, and robustness.
4. We propose a corresponding TA technique to mitigate distribution shift IL often suffers.

2 Imitation Game

We aim to cast text editing into an imitation game by defining the task as a recurrent sequence gen-
eration, as presented in Figure 2 (a). In this section, we describe the major components of our proposal,
including (1) the problem definition, (2) the data translation, (3) the model structure, and (4) a solution to the distribution shift.

\textsuperscript{1}Code and data are publicly available at GitHub.
2.1 Behavior cloning

We tear a text editing task \( \mathcal{X} \rightarrow \mathcal{Y} \) into recurrent subtasks of sequence generation \( S \rightarrow A \) defined by an MDP tuple \( \mathcal{M} = (S, A, P, \mathcal{E}, \mathcal{R}) \).

**State** \( S \) is a set of text sequences \( s = s_{1 \leq m}, \) where \( s \in \mathcal{V}_{S} \). We think of a source sequence \( \mathbf{x} \in \mathcal{X} \) as the initial state \( s_{1} \), its target sequence \( \mathbf{y} \in \mathcal{Y} \) as the goal state \( s_{T} \), and every edited sequence in between as an intermediate state \( s_{i} \). The path \( \mathbf{x} \rightarrow \mathbf{y} \) can be represented as a set of sequential states \( s_{1 \leq T} \).

**Action** \( A \) is a set of action sequences \( a = a_{i \leq m}, \) where \( a \in \mathcal{V}_{A} \). In Figure 3, “INSERT”, “POS_3”, and “=” are three action tokens belonging to the vocabulary space of action \( \mathcal{V}_{A} \). In contrast to token-level actions in sequence tagging, sentence-level ones set free the editing by varying edit metrics \( \mathcal{E} \) (e.g., Levenshtein distance) as long as \( \mathcal{X} \xrightarrow{\mathcal{E}} \mathcal{Y} \). It serves as an expert policy \( \pi^{*} \) to demonstrate the path to the goal state. A better expert usually means better demonstrations and imitation results. Hence, depending on the task, a suitable \( \mathcal{E} \) is essential.

**Transition matrix** \( P \) models the probability \( p \) that an action \( a_{i} \) leads a state \( s_{i} \) to the state \( s_{i+1} \). We know \( \forall s, a, p(s_{i+1}|s_{i}, a_{i}) = 1 \) due to the nature of text editing. So we can omit \( P \).

**Environment** \( \mathcal{E} \) responds to an action and updates the game state accordingly by \( s_{i+1} = \mathcal{E}(s_{i}, a_{i}) \) with process control. For example, the environment can refuse to execute actions that fail to pass the verification and terminate the game if a maximum number of iterations has been consumed.

**Reward function** \( \mathcal{R} \) calculates a reward for each action. It is a major factor contributing to the success of reinforcement learning. In the scope of this paper, we focus on BC, the simplest form of IL. So we can also omit \( \mathcal{R} \) and leave it for future work.

The formulation turns out to be a simplified \( \mathcal{M}_{BC} = (S, A, \mathcal{E}) \). Interacting with the environment \( \mathcal{E} \), we hope a trained agent is able to follow its learned policy \( \pi : S \rightarrow A \), and iteratively edit the initial state \( s_{0} = \mathbf{x} \) into the goal state \( s_{T} = \mathbf{y} \).

2.2 Trajectory generation

A data set to learn \( \mathcal{X} \rightarrow \mathcal{Y} \) consists of input-output pairs. It is necessary to convert it into state-action ones so that an agent can mimic the expert policy \( \pi^{*} : S \rightarrow A \) via supervised learning. A detailed TG is described in Algorithm 1.

Treating a pre-defined edit metric \( \mathcal{E} \) as the expert policy \( \pi^{*} \), we can leverage DP to efficiently find the minimum operations required to convert \( \mathbf{x} \) into \( \mathbf{y} \) in a left-to-right manner and backtrace this path to get specific operations.

Operations are later expressed as a set of sequential actions \( a_{i \leq T}^{*} \). Here we utilize a special symbol \( \text{DONE} \) to mark the last action \( a_{T}^{*} \), where \( \forall a \in a_{T}^{*}, a = \text{DONE} \). Once an agent performs \( a_{T}^{*} \), the current state is returned by the environment as the final output.

Given \( s_{1} = \mathbf{x} \), we attain the next state \( s_{2}^{*} = \)
\[ \mathcal{E}(s^*_i, a^*_i) \] and continue the rest until achieving \( s^*_T = y \), resulting in a set of sequential states \( s^*_1 \leq T \).

After one-to-one correspondence between states and actions, we collect a set of sequential expert’s demonstrations \( \tau^* = [(s^*_1, a^*_1), \ldots, (s^*_T, a^*_T)] \). Repeating the same process, we eventually convert \( \mathcal{X} \mapsto \mathcal{Y} \) into trajectories \( \mathcal{T}^* : \mathcal{S} \mapsto \mathcal{A} \).

### 2.3 Model architecture

We form \( \mathcal{S} \mapsto \mathcal{A} \) as sequence generation. More precisely, a neural model (i.e., the agent) takes states as input and outputs actions. Training an imitation model with BC corresponds to fitting a parametric model \( \pi_\theta \) that minimizes the negative log-likelihood loss \( l(a^*, \pi_\theta(s)) \). Most seq2seq models have an encoder-decoder structure.

**Encoder** takes an embedded state \( E(s) \in \mathbb{R}^{m \times d} \) and generates an encoded hidden state \( h_E \in \mathbb{R}^{n \times d} \) with \( d \) being the hidden dimension.

**Autoregressive decoder** in Figure 3 (a) conditions the current step on the encoded context and previously predictions to overcome the mismatch of sequence length. It calculates step by step

\[
 h_D^i = \text{AR}(E(a_{<i}), h_E) \in \mathbb{R}^d, \quad i = 0, \ldots, n + 1,
 \]

\[
 \hat{a}_i = \text{LogSoftmax}(h_D^i) \in \mathbb{R}^{|V_A|}, \quad i = 0, \ldots, n + 1,
\]

and in the end, returns \( \hat{a} \in \mathbb{R}^{n \times |V_A|} \). The training is conducted as back-propagating \( l(a^*, \hat{a}) \). Note that \( a_0^0 = \text{BOS} \) and \( a_{n+1}^0 = \text{EOS} \) encourage the decoder to learn to begin and end the autoregression.

**Non-autoregressive decoder** instead provides hidden states in one time. It is feasible to apply techniques of non-autoregressive machine translation. However, one of the primary issues solved by that is the uncertainty of the target sequence length. When it comes to state-action prediction, thanks to the flexibility at the sequence-level, we are allowed to design actions on purpose to eliminate such uncertainty. Specifically, we enforce action sequences to be of fixed length. On this basis, we propose D2 as shown in Figure 3 (b). To address the misalignment of sequence length between state and action, we insert a fully connected feed-forward network between the encoder and decoder_0.

\[
 \text{FFN}(h_E) = (h_E^T W + b)^T \in \mathbb{R}^{n \times d}
\]

where \( W \in \mathbb{R}^{m \times n} \) and \( b \in \mathbb{R}^{d \times n} \) transform the length dimension from \( m \) to \( n \) so as to project \( h_E \) into \( h_F \in \mathbb{R}^{n \times d} \). The alignment of the sequence length allows us to trivially pass \( h_F \) to decoder_0.

\[
 h_{D_0} = \text{NAR}_0(h_F, h_E) \in \mathbb{R}^{n \times d}
\]

\[
 \hat{a}^0 = \text{LogSoftmax}(h_{D_0}) \in \mathbb{R}^{n \times |V_A|}
\]

For a clear comparison with the autoregressive decoder, we make minimal changes to the structure and keep modeling the dependence between two contiguous steps through decoder_1. To elaborate, we shift \( \hat{a}^0 \) one position to the right as \( \hat{a}^1 \) by appending \( a_0^0 \) at the beginning and remove \( a_{n+1}^0 \) to maintain the sequence length. After that, we continue to feed \( \hat{a}^0 \) to decoder_1.

\[
 h_{D_1} = \text{NAR}_1(h_D^0, h_E) \in \mathbb{R}^{n \times d}
\]

\[
 \hat{a}^1 = \text{LogSoftmax}(h_{D_1}) \in \mathbb{R}^{n \times |V_A|}
\]

At last, we conduct backpropagation with respect to the loss summation \( l(a^*, \hat{a}^0) \oplus l(a^*, \hat{a}^1) \). Conventional seq2seq architectures are often equipped with intermediate modules such as a full attention distribution over the encoded context (Bahdanau et al., 2015), which is omitted in the above formulation for simplicity. In the implementation, we always assume to train decoder_0 and decoder_1 separately to increase the model capacity, yet weight sharing is possible.
2.4 Trajectory augmentation

IL suffers from distribution shift and error accumulation (Ross et al., 2011). An agent’s mistakes can easily put it into a state that the expert demonstrations do not involve and the agent has never seen during training. This also means errors can add up, so the agent drifts farther and farther away from the demonstrations. To tackle this issue, we propose TA that expands the expert demonstrations and actively exposes shifted states to the agent. We accomplish this by diverting intermediate states and consider them as initial states for TG. An example is offered in Figure 2 (b).

Given expert states $s_{t<T}$ and corresponding actions $a_{t<T}$, we utilize the divide-and-conquer technique to (1) break down the chain of state generation $s_{t} \rightarrow a_{t} \rightarrow s_{t+1}$ into two by either executing $a_{t}$ to stay on the current path or skipping $a_{t}$ to branch the current path; (2) recursively calling this process until reaching the goal state $s_{T}^{∗}$; (3) merge intermediate states from branches and return from bottom to top in the end. As illustrated in Algorithm 2, we collect a set of shifted states

$$S’ = TA(\emptyset, s_{t}^{∗}, s_{t<T}, a_{t<T}, \mathcal{E}),$$

regard them as initial states paired with the same goal state to produce extra trajectories

$$\tau’ = \bigcup_{s^{∗}\in S’} TG(s^{∗}, s_{t}^{∗}, \mathcal{E}, \mathcal{E}),$$

and finally yield the augmented expert demonstrations $\mathcal{T}’ \cup \mathcal{T}’$ after looping through $\mathcal{X}$.

TA is advantageous because it (i) only exploits existing expert demonstrations to preserve the i.i.d assumption; (ii) is universally applicable to our proposed paradigm without a dependency on the downstream task; (iii) does not need domain knowledge, labeling work, and further evaluation.

3 Experiments

We adapt recurrent inference to our paradigm and evaluate them across AE benchmarks.

3.1 Setup

Data. Arithmetic Operators Restoration (AOR) is a short-to-long editing to complete an array into a true equation. It is also a one-to-many task as an array can be completed as multiple true equations differently. Arithmetic Equation Simplification (AES) aims to calculate the parenthesized parts and keep the equation hold, resulting in a long-to-short and many-to-one editing. Arithmetic Equation Correction (AEC) targets to correct potential mistakes in an equation. Diverse errors perturb the equation, making AEC a mixed many-to-many editing. To align with the previous work, we follow the same data settings $N$, $L$, and $D$ for data generation, as well as the same action design for trajectory generation. The edit metric $E$ for AOR and AEC is Levenshtein, while $E$ for AES is a self-designed one (SELF) that instructs to replace tokens between two parentheses with the target token. Examples are presented in Table 2. We refer readers to Shi et al. (2020) for an exhaustive explanation. As shown in Table 1, the data splits are 7K/1.5K/1.5K for training, validation, and testing respectively.

Evaluation. Sequence accuracy and equation accuracy are two primary metrics with token accuracy for a more fine-grained reference. In contrast to sequence accuracy for measuring whether an equation exactly matches the given label, equation accuracy emphasizes whether an equation holds, which is the actual goal of AE tasks. It is noted that there is no hard constraint to guarantee that all the predicted actions are valid. However, when the agent makes an inference mistake, the environment can refuse to execute invalid actions and keep the current state. This is also one of the beauties of reformulating text editing as a controllable MDP.

3.2 Baselines. Recurrent inference (Recurrence) exhibits advantages over conventional end-to-end (End2End) and sequence tagging (Tagging) (Shi et al., 2020). However, for AES and AEC, it allows feeding training samples to a data generator and exposing more variants to models. These variants, as source samples paired with corresponding target samples, are used as the augmented dataset. This is impractical due to the strong dependency on domain knowledge. Given an input $1 + (2 + 2) = 2$.
5" and output "1 + 4 = 5" in AES, a variant "1 + (1 + 3) = 5" can be generated based on the knowledge 1 + 3 = 4. Nevertheless, if this knowledge is not provided in the other training samples, the model should only know 2 + 2 = 4.

**Models.** As discussed, since the previously reported experiments are not practical, we re-run Recurrence source code for a more reasonable baseline (Recurrence*) that only has access to the fixed training set. Meanwhile, in our development environment, we reproduce Recurrence* within the proposed paradigm according to the compatibility in between. The encoder-decoder architecture inherits the same recurrent network as the backbone with long short-term memory units (Hochreiter and Schmidhuber, 1997) and an attention mechanism (Luong et al., 2015). The dimension of the bidirectional encoder is 256 in each direction and 512 for both the embedding layer and decoder. We apply a dropout of 0.5 to the output of each layer (Srivastava et al., 2014). This provides us a standard autoregressive baseline AR, as well as a more powerful AR* after increasing the number of encoder layers from 1 to 4. On the one hand, to construct a non-autoregressive baseline NAR, we replace the decoder of AR* with a linear layer that directly maps the context to a probability distribution over the action vocabulary. In addition, we add two more encoder layers to maintain a similar amount of trainable parameters. On the other hand, replacing the decoder of AR* with D2 leads to our model NAR*. We strictly unify the encoder for a fair comparison regarding the decoder. Model configurations are shared across AE tasks for a comprehensive assessment avoiding particular tuning against any of them.

**Training.** We train on a single NVIDIA Titan RTX with a batch size of 256. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $10^{-3}$ and an $\ell_2$ gradient clipping of 5.0 (Pascaru et al., 2013). A cosine annealing scheduler helps manage the training process and restarts the learning every 32 epochs to get it out of a potential local optimum. We adopt early stopping to wait for a lower validation loss until there are no updates for 512 epochs (Prechelt, 1998). Teacher forcing with a rate of 0.5 spurs up the training process (Williams and Zipser, 1989). In AES and AEC, the adaptive loss weighting guides the model to adaptively focus on particular action tokens in accordance with the training results. Reported metrics attached with standard deviation are the results of five runs using random seeds from [0, 1, 2, 3, 4].

<table>
<thead>
<tr>
<th>Term</th>
<th>Source $x$</th>
<th>Target $y$</th>
<th>State $s_1^t$</th>
<th>Action $a_1^t$</th>
<th>Next State $s_{t+1}^t$</th>
<th>Shifted State $s_{t+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 + 3 = 4$</td>
<td>3 + 6 / 2 9 3</td>
<td>- 3 - 6 / 2 + 9 = 3</td>
<td>- 3 - 6 / 2 9 3</td>
<td>[POS_6, +]</td>
<td>- 3 - 6 / 2 + 9 3</td>
<td>- 3 - 6 / 2 9 3</td>
</tr>
<tr>
<td>$5 + 5 = 10$</td>
<td>5 + 5 + 96 = 3</td>
<td>65 + 5 + (64 + 32) = (- 25 + 58)</td>
<td>65 + 5 + (64 + 32) = (- 25 + 58)</td>
<td>[POS_4, POS_8, 96]</td>
<td>65 + 5 + 96 + (83 - 24) = (- 25 + 58)</td>
<td>65 + 5 + (64 + 32) + 59 = (- 25 + 58)</td>
</tr>
<tr>
<td>$5 - 2 = 3$</td>
<td>5 - (64 + 32) = (- 25 + 58)</td>
<td>- 3 + 6 / 2 9 3</td>
<td>65 + 5 + (64 + 32) = (- 25 + 58)</td>
<td>[DELETE, POS_3, POS_3]</td>
<td>65 + 5 + (64 + 32) + 59 = (- 25 + 58)</td>
<td>65 + 5 + (64 + 32) + 59 = (- 25 + 58)</td>
</tr>
</tbody>
</table>

Table 1: Data statistics of AE benchmarks.

Table 2: Examples from AE with specific $N$ for integer size, $L$ for the number of integers, and $D$ for data size.
Table 3: Evaluation results on AOR, AES, and AEC with specific $N$, $L$, and $D$. The token and sequence accuracy for AOR were not reported, thus we leave these blank here. With or without TA, our proposed NAR* achieves the best performance in terms of equation accuracy across the board.

Figure 4: The learning curve of AR* (left column) and NAR* (right column) across AE tasks (rows). The red and blue lines represent the training on actions w.r.t. sequence accuracy. The orange line stands for the validation on returned states w.r.t. equation accuracy. The dashed line in green marks the earlier stop epoch of NAR* than that of AR* during training.

Non-autoregressive. What stands out is the dominance of NAR*, achieving 61.45%, 95.67%, and 46.43% in terms of equation accuracy for AOR, AES, and AEC, respectively. Particularly in AES, its better performance over AR* by more than 2.1% equation accuracy underlines the success of NAR* in capturing the interdependencies among target tokens. Its superiority with respect to equation accuracy boosting by around 66.18% over NAR highlights the contributions of D2 again.

Trajectory augmentation. As expected, the incorporation of TA consistently promotes the accuracy of all models in our learning regime throughout AE tasks. Taking NAR as an example, training with TA brings it a substantial equation accuracy gain, remarkably up to 18.42% in AES. Even more, it pushes the gap between NAR* and the other baselines. The most notable advance comes from AOR, where NAR* outperforms AR* by a substantial margin of 5.68% equation accuracy. It appears that TA is more effective for non-autoregressive models than autoregressive ones.

4 Analysis

We conduct extensive sensitivity analyses to better illustrate and understand our methods.

4.1 Efficiency

From the learning curve (Figure 4) and inference time (Figure 5) of AR* and NAR* in AE, in addition to a higher accuracy, we find NAR* needs less number of training epochs to converge and trigger the early stopping. The periodic fluctuation of the learning curve is the consequence of using a scheduler. When it comes to inference, NAR* saves much time for every step of action determin-
<table>
<thead>
<tr>
<th>Design</th>
<th>Action Sequence</th>
<th>Method</th>
<th>Tok. Acc. %</th>
<th>Eq. Acc. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>[Pos. L, Pos. R, Tok.]</td>
<td>AR*</td>
<td>99.27±0.32</td>
<td>93.57±2.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAR*</td>
<td>99.51±0.13</td>
<td>95.67±0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR*+TA</td>
<td>99.44±0.27</td>
<td>95.24±2.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAR*+TA</td>
<td>99.58±0.15</td>
<td>96.44±1.29</td>
</tr>
<tr>
<td>#2</td>
<td>[Pos. L, Tok., Pos. R]</td>
<td>AR*</td>
<td>99.08±0.93</td>
<td>92.35±4.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAR*</td>
<td>99.50±0.27</td>
<td>95.55±2.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR*+TA</td>
<td>99.52±0.29</td>
<td>96.68±2.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAR*+TA</td>
<td>99.54±0.20</td>
<td>95.97±1.64</td>
</tr>
<tr>
<td>#3</td>
<td>[Tok., Pos. L, Pos. R]</td>
<td>AR*</td>
<td>99.06±0.79</td>
<td>83.79±6.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAR*</td>
<td>99.53±0.14</td>
<td>95.99±0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR*+TA</td>
<td>99.41±0.49</td>
<td>87.29±3.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAR*+TA</td>
<td>99.61±0.06</td>
<td>96.55±0.46</td>
</tr>
</tbody>
</table>

Table 4: Evaluation of AR* and NAR* in AES across three action designs that vary from each other by token order. They direct to the same operation with Pos_L/Pos_R/Tok., denoting left parenthesis/right parenthesis/target token.

nation and ends up returning the edited state faster. As AR* and NAR* share exactly the same encoder structure, we conclude that D2 contributes to the advanced efficiency.

4.2 Action design

Due to the liberty of sequence generation, the same operation can be represented as different action sequences. In AES, the operation, instructing to substitute tokens between left and right parentheses with the required token, can fit the three action designs in Table 4, where Pos_L, Pos_R, and Tok. denote the positions of two parentheses and the target token. Design #1 is the default one. A simple swap of action tokens offers designs #2 and #3.

AR* severely suffers such perturbation, causing an equation accuracy decline by 9.78% in #3. Contrastly, NAR* holds around its results and even slightly improves to 95.99% in #3. Despite the joining of TA, AR* still goes down from 95.24% in #1 to 87.29% in #3, while NAR* stays nearly consistent across three designs. It is reasonable that AR* is sensitive to the order of action tokens because the position information helps the inference of the target token. This also reflects that NAR* can catch the position information but with little dependence on token order. Such robustness allows greater freedom of action design.

4.3 Trajectory optimization

A better edit metric E often means a smaller action vocabulary space $|\mathcal{V}_A|$, shorter trajectory length $T_{\text{max}}$, and, therefore, an easier IL. Taking AES as an instance, a SELF-action, replacing tokens enclosed in parentheses with the target one, actually is the compression of several Levenshtein-actions including multiple deletions and one substitution. Although either can serve as an expert policy, SELF causes a much shorter $T_{\text{max}}$ as indicated in Table 5. The change from SELF to Levenshtein brings on a longer $T_{\text{max}}$ and consequently a significant performance gap of 75.2% and 77.74% for AR* and NAR* in terms of equation accuracy. Doing one edit in 31 steps rather than 6 undoubtedly raises the difficulty of the imitation game.

As one more exploration, we introduce Longest Common Subsequence (LCS) as an alternative E to AES. Token replacement is not allowed in LCS but in Levenshtein. A replacement action has to be decomposed as one deletion and one insertion in LCS. From this, LCS has a small $|\mathcal{V}_A|$, while Levenshtein has a shorter $T_{\text{max}}$. We train NAR* with these two and report in Figure 6. For a clear comparison, the test set is divided into two groups. In w/o REPLACE, both yield the same $T_{\text{max}}$, but, in w/ REPLACE, Levenshtein takes a shorter $T_{\text{max}}$.
As an ablation study, we freeze the encoder of D2 without TA. In the latter, the opposite is true, where Table 6, replacing the decoder with a linear layer distribution shift alleviation. Achievements on coupled with TG for data translation and TA for autoregressive decoder for state-action learning, 5 Conclusion

In the former, LCS exceeds Levenshtein with or without TA. In the latter, the opposite is true, where Levenshtein outperforms LCS under the same condition. This support our assumption at the beginning that an appropriate E, leading to a small |ṼA| and a short Tmax, is conducive to IL, suggesting trajectory optimization an interesting future work.

4.4 Dual decoders

As an ablation study, we freeze the encoder of NAR* and vary its decoder to reveal the contributions of each component in D2. As listed in Table 6, replacing the decoder with a linear layer leads to Linear and removing the second decoder from NAR* results in Decoder0. Moreover, sharing the parameters between two decoders of NAR* gives the Shared Dec2. All of them can parallel the decoding process. We then borrow the setup of Section 3 and test them on AE.

Among four decoders, NAR* dominates three imitation games. The performance decrease caused by shared parameters is more significant than expected. Besides the reason that saved parameters limit the model capacity, another potential one is the input mismatch of two decoders. The output of decoder0 is the projected context from the linear layer after the encoder, yet that of decoder1 is the embedded prediction from the embedding layer. When incorporating TA, we find the same trend persists. The gap between NAR* and the others is even more apparent. Since they share the same encoder, such a gap clarifies the benefits of D2.

5 Conclusion

We reformulate text editing as an imitation game defined by an MDP to allow action design at the sequence-level. We propose D2, a non-autoregressive decoder for state-action learning, coupled with TG for data translation and TA for distribution shift alleviation. Achievements on AE benchmarks evidence the advantages of our methods in performance, efficiency, and robustness. Sequence-level actions are arguably more controllable, interpretable, and similar to human behavior. Turning tasks into games that agents feel more comfortable with sheds light on future studies in the direction of reinforcement learning in the application of text editing. The involvement of a reward function, the optimization of the trajectories, the design of sequence-level actions, and their applications in more practical tasks, to name a few, are interesting for future work. Suggesting text editing as a new testbed, we hope our findings will shed light on future studies in reinforcement learning applying to natural language processing.

Limitations

Each time the state is updated, the agent can get immediate feedback on the previous action and thus a dynamic context representation during the editing. This also means that the encoder (e.g., a heavy pretrained language model) will be called multiple times to refresh the context matrix. Consequently, as the trajectory grows, the whole task becomes slow even though we have paralleled the decoding process. Meanwhile, applying our methods in more realistic editing tasks (e.g., grammatical error correction) remains a concern and needs to be explored in the near future.

Acknowledgements

We gratefully appreciate Che Wang (Watcher), Yichen Gong, and Hui Xue for sharing their pearls of wisdom. We also would like to express our special thanks of gratitude to Yingying Huo for the support, as well as EMNLP anonymous reviewers for their constructive feedback. This work was supported by Shining Lab and Alibaba Group.
References


