Appendix: Non-Monotonic Sequential Text Generation

Sean Welleck 1  Kianté Brantley 2  Hal Daumé III 2 3  Kyunghyun Cho 1 4 5

A. Additional Experiment Details and Results

A.1. Word Reordering

Model The decoder is a 2-layer LSTM with 1024 hidden units, dropout of 0.0, based on a preliminary grid search of $n_{\text{layers}} \in \{1, 2\}$, $n_{\text{hidden}} \in \{512, 1024, 2048\}$, dropout $\in \{0.0, 0.2, 0.5\}$. Word embeddings are initialized with GloVe vectors and updated during training. All presented Word Reordering results use greedy decoding.

Training Each model was trained on a single GPU using a maximum of 500 epochs, batch size of 32, Adam optimizer, gradient clipping with maximum $\ell_2$-norm of 1.0, and a learning rate starting at 0.001 and multiplied by a factor of 0.5 every 20 epochs. For evaluation we select the model state which had the highest validation BLEU score, which is evaluated after each training epoch.

Oracle For $\pi^*_{\text{annealed}}$, $\beta$ is linearly annealed from 1.0 to 0.0 at a rate of 0.05 each epoch, after a burn-in period of 20 epochs and multiplied by a factor of 0.5 every 20 epochs. For evaluation we select the model state which had the highest validation BLEU score, which is evaluated after each training epoch.

Example Predictions Figure 4 shows example predictions from the validation set, including the generation order and underlying tree.

A.2. Unconditional Generation

We use the same settings as the Word Reordering experiments, except we always use stochastic sampling from $\pi^*_{\text{coaching}}$ during roll-in. For evaluation we select the model state at the end of training.

Oracle For $\pi^*_{\text{annealed}}$, $\beta$ is linearly annealed from 1.0 to 0.0 at a rate of 0.05 each epoch, after a burn-in period of 20 epochs and multiplied by a factor of 0.5 every 20 epochs. We do not observe significant performance variations with stochastically sampling from $\pi^*_{\text{coaching}}$. These settings are based on a grid search of $\beta_{\text{rate}} \in \{0.01, 0.05\}$, $\beta_{\text{burn-in}} \in \{0, 20\}$, coaching-rollin $\in \{\text{greedy, stochastic}\}$ using the model selected in the Model section above.

Example Predictions Figure 4 shows example predictions from the validation set, including the generation order and underlying tree.

A.3. Machine Translation

Data and Preprocessing We use the default Moses tokenizer script (Koehn et al., 2007) and segment each word into a subword using BPE (Sennrich et al., 2015) creating 40k tokens for both source and target. Similar to (Bahdanau et al., 2015), during training we filter sentence pairs that exceed 50 words.

Transformer Policy The Transformer policy uses 4 layers, 4 attention heads, hidden dimension 256, feed-forward dimension 1024, and is trained with batch-size 32 and a

---

Table 1. Unconditional generation BLEU for various top-k samplers and policies trained with the specified oracle.

<table>
<thead>
<tr>
<th>Oracle</th>
<th>$k$</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi^*_{\text{left-right}}$</td>
<td>10</td>
<td>0.905</td>
<td>0.778</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.874</td>
<td>0.705</td>
<td>0.514</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.853</td>
<td>0.665</td>
<td>0.466</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>0.853</td>
<td>0.668</td>
<td>0.477</td>
</tr>
<tr>
<td>$\pi^*_{\text{uniform}}$</td>
<td>10</td>
<td>0.966</td>
<td>0.906</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.916</td>
<td>0.751</td>
<td>0.544</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.864</td>
<td>0.651</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>0.831</td>
<td>0.609</td>
<td>0.395</td>
</tr>
<tr>
<td>$\pi^*_{\text{annealed}}$</td>
<td>10</td>
<td>0.966</td>
<td>0.895</td>
<td>0.770</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.931</td>
<td>0.804</td>
<td>0.628</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.907</td>
<td>0.765</td>
<td>0.585</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>0.894</td>
<td>0.740</td>
<td>0.549</td>
</tr>
</tbody>
</table>

Unconditional Samples Samples in Tables 3-4 are organized as ‘short’ ($\leq$ 5th percentile), ‘average-length’ (45-55th percentile), and ‘multi-sentence’ ($\geq$ 3 punctuation tokens). Each image in Figures 1, 2, and 3 shows a sampled sentence, its underlying tree, and its generation order.

Additional BLEU Scores Since absolute BLEU scores can vary by using a softmax temperature (Caccia et al., 2018) or top-k sampler, we report additional scores for $k \in \{10, 100, 1000\}$ and BLEU-$\{2, 3, 4\}$ in Table 1. Generally the policy trained with the annealed oracle achieves the highest metrics.
Appendix: Non-Monotonic Sequential Text Generation

<table>
<thead>
<tr>
<th>Oracle</th>
<th>BLEU (BP)</th>
<th>Meteor</th>
<th>YiSi</th>
<th>Ribes</th>
<th>BLEU (BP)</th>
<th>Meteor</th>
<th>YiSi</th>
<th>Ribes</th>
</tr>
</thead>
<tbody>
<tr>
<td>left-right</td>
<td>29.47 (0.97)</td>
<td>29.66</td>
<td>52.03</td>
<td>82.55</td>
<td>26.23 (0.97)</td>
<td>27.87</td>
<td>47.58</td>
<td>79.85</td>
</tr>
<tr>
<td>uniform</td>
<td>14.97 (0.63)</td>
<td>21.76</td>
<td>41.62</td>
<td>77.70</td>
<td>13.17 (0.64)</td>
<td>19.87</td>
<td>36.48</td>
<td>75.36</td>
</tr>
<tr>
<td>+⟨end⟩-tuning</td>
<td>18.79 (0.89)</td>
<td>25.30</td>
<td>46.23</td>
<td>78.49</td>
<td>17.68 (0.96)</td>
<td>24.53</td>
<td>42.46</td>
<td>74.12</td>
</tr>
<tr>
<td>annealed</td>
<td>19.50 (0.71)</td>
<td>26.57</td>
<td>48.00</td>
<td>81.48</td>
<td>16.94 (0.72)</td>
<td>23.15</td>
<td>42.39</td>
<td>78.99</td>
</tr>
<tr>
<td>+⟨end⟩-tuning</td>
<td>21.95 (0.90)</td>
<td>26.74</td>
<td>49.01</td>
<td>81.77</td>
<td>19.19 (0.91)</td>
<td>25.24</td>
<td>43.98</td>
<td>79.24</td>
</tr>
</tbody>
</table>

Table 2. LSTM Policy results for machine translation experiments.

learning rate $10^{-5}$. For this model and experiment, we define an epoch as 1,000 model updates. The learning rate is divided by a factor of 1.1 every 100 epochs. For $\pi^\star_{\text{annealed}}$, $\beta$ is linearly annealed from 1.0 to 0.0 at a rate of 0.01 each epoch, after a burn-in period of 100 epochs. We compute metrics after each validation epoch, and following training we select the model with the highest validation BLEU.

**Loss with Auxiliary ⟨end⟩ Predictor**  A binary cross-entropy loss is used for the ⟨end⟩ predictor for all time-steps, so that the total loss is $L_{\text{bce}}(\pi^\star, \pi_{\text{end}}) + L_{\text{KL}}(\pi^\star, \pi)$ where $L_{\text{KL}}$ is the loss from Section 3.2. For time-steps in which ⟨end⟩ is sampled, $L_{\text{KL}}$ is masked, since the policy’s token distribution is not used when $a_t$ is ⟨end⟩. $L_{\text{KL}}$ is averaged over time by summing the loss from unmasked time-steps, then dividing by the number of unmasked time-steps.

**Tree Position Encodings**  We use an additional tree position encoding, based on (Shiv & Quirk, 2019), which may make it easier for the policy to identify and exploit structural relationships in the partially decoded tree. Each node is encoded using its path from the root, namely a sequence of left or right steps from parent to child. Each step is represented as a 2-dimensional binary vector ([0, 0] for the root, [1, 0] for left and [0, 1] for right), so that the path is a vector $e(a_i) \in \{0, 1\}^{2^{\text{max-depth}}}$ after zero-padding. Finally, $e(a_i)$ is multiplied element-wise by a geometric series of a learned parameter $p$, that is, $e(a_i) \cdot [1, p, p^2, p^3, ...]$. We only use this approach with the Transformer policy.

**Additional LSTM Policy**  Results are shown in Table 2. We use a bi-directional LSTM encoder-decoder architecture that has a single layer of size 512, with global concat attention (Luong et al., 2015). The learning rate is initialized to 0.001 and multiplied by a factor of 0.5 on a fixed interval.

**References**


### Table 3. Short (left) and Average-Length (right) unconditional samples from policies trained on Persona-Chat.

<table>
<thead>
<tr>
<th>left-right</th>
<th>right</th>
<th>uniform</th>
<th>right</th>
<th>annealed</th>
<th>right</th>
</tr>
</thead>
<tbody>
<tr>
<td>i can drive you alone.</td>
<td>do you like to test your voice to a choir?</td>
<td>i am freelance a writer but i am a writer.</td>
<td>i love meat. or junk food. i sometimes go too much i make. avoid me unhealthy.</td>
<td>i am definitely a kid. are you? i am 10!</td>
<td>i am in michigan state... that is a grand state.</td>
</tr>
<tr>
<td>yeah it is very important.</td>
<td>no pets, on the subject in my family, yes.</td>
<td>that is so sad. do you have a free time?</td>
<td>does not kill anyone that can work around a lot of animals? you? i like trains.</td>
<td>i am well. thank you. i love my sci fi stories. i write books.</td>
<td>that is good. i work as a pharmacist in florida...</td>
</tr>
<tr>
<td>i am a am nurse.</td>
<td>cool. i have is also a cat named cow.</td>
<td>yes i do not like pizza which is amazing lol.</td>
<td>baby? it will it all here. that is the workforce.</td>
<td>i am good. thank you. i love my sci fi stories. i write books.</td>
<td>how are you? wanna live in san fran! i love it.</td>
</tr>
<tr>
<td>do you actually enjoy it?</td>
<td>i am doing good taking a break from working on it.</td>
<td>since the gym did not bother me many years ago.</td>
<td>that is interesting. i am just practicing my piano degree.</td>
<td>i am well. thank you. my little jasper is new.</td>
<td>well that is awesome! i do crosswords! that is cool.</td>
</tr>
</tbody>
</table>

### Table 4. Multi-sentence unconditional samples from policies trained on Persona-Chat.

<table>
<thead>
<tr>
<th>left-right</th>
<th>right</th>
<th>uniform</th>
<th>right</th>
<th>annealed</th>
<th>right</th>
</tr>
</thead>
<tbody>
<tr>
<td>nice! i think i will get a jump blade again. have you done that at it?</td>
<td>just that is for a while. and yourself right now?</td>
<td>i like to be talented.</td>
<td>yeah it can be. what is your favorite color?</td>
<td>i am definitely a kid. are you? i am 10!</td>
<td>i am in michigan state... that is a grand state.</td>
</tr>
<tr>
<td>great. what kinds of food do you like best? i love italian food.</td>
<td>i am freelance a writer but i am a writer.</td>
<td>how are you doing buddy?</td>
<td>i do not have dogs. they love me here.</td>
<td>i am well. thank you. i love my sci fi stories. i write books.</td>
<td>that is good. i work as a pharmacist in florida...</td>
</tr>
<tr>
<td>wow. bike ride is my thing. i do nothing for kids.</td>
<td>that is so sad. do you have a free time?</td>
<td>i like healthy foods.</td>
<td>no kids... i am... you?</td>
<td>i am well. thank you. my little jasper is new.</td>
<td>how are you? wanna live in san fran! i love it.</td>
</tr>
<tr>
<td>i am alright. my mom makes work and work as a nurse. that is what i do for work.</td>
<td>yes i do not like pizza which is amazing lol.</td>
<td>i love to eat.</td>
<td>that is interesting. i am just practicing my piano degree.</td>
<td>i am well. thank you. my little jasper is new.</td>
<td>well that is awesome! i do crosswords! that is cool.</td>
</tr>
</tbody>
</table>
Figure 1. Unconditional samples from a policy trained with $\pi^\text{annealed}$.

Sentence: I wish you could study lol. I work a lot.
Gen. Order: you, I study I wish could lol a work lot

Sentence: I work from home. Just hoping to find home at music, hoping you will do it.
Gen. Order: I work, home music you from home hoping will just at do hoping it find to

Sentence: I do. I like lipton beverages.
Gen. Order: I do I like beverages lipton

Sentence: I love some of my own green gables.
Gen. Order: I some love my of own gables green
Figure 2. Unconditional samples from a policy trained with $\pi_{\text{uniform}}$.

Sentence: yeah first patrol which i want to touch and teach the first ? i am math d .
Gen. Order: want yeah touch first to teach which math patrol i the . i math first am

Sentence: hi there , back a football game ? i got the bench track , could you feel so short .
Gen. Order: the ? could a get . short , game i track you . there back football bench feel hi so

Sentence: eh really ? that ? it is dangerous !
Gen. Order: is ? dangerous eh that ! really it ?

Sentence: long hair ! i am recovering from it in programming .
Gen. Order: am i . hair from long ! recovering in it programming

Sentence: does it involve a family , or anyone known anyone because i can dance if i tell ?
Gen. Order: known family can it , i dance does a anyone ? involve or because if i tell

Sentence: thanks . what favorite kind of car does your make fast food .
Gen. Order: of favorite your thanks kind does food . car fast . what make

Sentence: both do not think so too ! many different in georgia or ?
Gen. Order: in both ? many or ! different georgia do too so think not

Sentence: i am running , walking and chasing cheetahs in the park .
Gen. Order: . and walking cheetahs , chasing the i in park running an
Figure 3. Unconditional samples from a policy trained with $\pi_{\text{left-right}}$.

Sentence: fancy. do you have any hobbies?
Gen. Order: fancy. do you have any hobbies?

Sentence: a little sick, but i am thinking about a new diet.
Gen. Order: a little sick, but i am thinking about a new diet.

Sentence: she won more of work on yourself. made life down.
Gen. Order: she won more of work on yourself. made life down.

Sentence: i do not wish i were a better person.
Gen. Order: i do not wish i were a better person.

Sentence: mostly classical. it is very good in our lives.
Gen. Order: mostly classical. it is very good in our lives.

Sentence: over being a lumberjack. i love to be outdoors. its just me and listening to music.
Gen. Order: over being a lumberjack. i love to be outdoors. its just me and listening to music.

Sentence: oh yeah that happens! i always check the watch out.
Gen. Order: oh yeah that happens! i always check the watch out.

Sentence: neat when i visit art school. i hope i love them.
Gen. Order: neat when i visit art school. i hope i love them.
Figure 4. Word Reordering Examples. The columns show policies trained with $\pi_{\text{left-right}}$, $\pi_{\text{uniform}}$, and $\pi_{\text{annealed}}$, respectively.

Actual: "I like to live outside but I would have never been of this world." Predicted: "I like to live outside but I would have never been of this world.

Actual: "I am retired for now and been loving it." Predicted: "I am retired for now and been loving it."
Figure 5. Translation outputs from a policy trained with $\pi_{annealed}$ on the test set.

Source: Siehe die Aktivitäten der Tierschutzorganisation.
Target: Die Tierschutzorganisation betreibt einen Wildschutz.
Gen. Order: Sie sehen, wie die Tierschutzorganisation ihre Arbeit betreibt.

Source: Der Chef sagte mir, dass wir in zwei Wochen endlich starten werden.
Target: Der Chef sagte mir, dass wir in zwei Wochen endlich starten werden.
Gen. Order: Der Chef sagte mir, dass wir in zwei Wochen endlich starten werden.

Source: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.
Target: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.

Source: Wie viele Leute sind in der Nähe?
Target: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.
Gen. Order: Wie viele Leute sind in der Nähe?

Source: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.
Target: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.

Source: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.
Target: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.

Source: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.
Target: Ich habe gelernt, dass die Aktivitäten der Tierschutzorganisation sehr wichtig sind.
Appendix: Non-Monotonic Sequential Text Generation

Figure 6. Translation outputs from a policy trained with $\pi_{\text{uniform}}$ on the test set.
Appendix: Non-Monotonic Sequential Text Generation

Figure 7. Translation outputs from a policy trained with \( \pi_{\text{train}} \) on the test set.

Source: mit gewissen Vorbehalten die Griechen auch der Violenkreis .
Target: so with this tool , this boundary has been broken .
Predicted: so this instrument was taken through that motor limit .
Gen. Order: so this instrument was taken through that motor limit .

Source: einige meiner vollkommenen Kinder schätzen nicht besonders gut es .
Target: some of my perfect kids aren ’ t doing so well .
Predicted: some of my smart kids don ’ t do it run very well .
Gen. Order: some of my smart kids don ’ t do it run very well .

Source: eine junge Unternehmen haben eine enorme Auswirkung auf die Wirtschaft .
Target: these young enterprises have a tremendous impact on their nation .
Predicted: these young companies have a tremendous impact on their nation .
Gen. Order: these young companies have a tremendous impact on their nation .

Source: wer wird mich verlassen , werde ich Knie habe ? Gött !
Target: who is going to take care of me if I have cancer ?
Predicted: who will feed me when I have cancer ?
Gen. Order: who will feed me when I have cancer ?

Source: einen einzigartiger Traum kann einen Zweck erfüllen .
Target: even a shattered dream can do that for you .
Predicted: even a shattered dream can meet this purpose .
Gen. Order: even a shattered dream can meet this purpose .

Source: since 1970 waren keine bekannten desasters mehr auf der mond .
Target: since 1970 , no human being has been back to the moon .
Predicted: since 1970 , no human being has been back to the moon .
Gen. Order: since 1970 , no human being has been back to the moon .