PAIR: Planning and Iterative Refinement in Pre-trained Transformers for Long Text Generation

Xinyu Hua¹ and Lu Wang²

¹Northeastern University
²University of Michigan
Motivation

• Large pre-trained Transformers have shown great promise for text generation applications.
Motivation

• Large pre-trained Transformers have shown great promise for text generation applications. [Roller+, 2020]

https://newsyoucantuse.com/
Motivation

• Auto-regressive seq2seq:

Encoder (bidirectional)

\[ x_1 \ x_2 \ x_3 \ldots \ x_{t-1} \ x_t \]

Decoder (Auto-regressive)

\[ \text{[BOS]} \ y_1 \ y_2 \ y_3 \ ? \]

Cross-attention
Motivation

• Auto-regressive seq2seq:

Encoder (bidirectional)

$x_1 \ x_2 \ x_3 \ldots \ x_{t-1} \ x_t$

Decoder (Auto-regressive)

[y_1 \ y_2 \ y_3 \ ?]

Cross-attention

GPT-2 (decoder only)
Major Challenges

• Lack of coherent text plans [Wiseman+, 2018; Moryossef+, 2019; Hua & Wang, 2019]

• Low controllability [Wen+, 2015; Hu+, 2017; Dathathri+, 2020; Goyal & Durrett, 2020]

• Hard to correct errors [Guu+, 2018; Lee+, 2018; Ghazvininejad+, 2019]
Motivation

- Masked language model:

Encoder (bidirectional)

Decoder (Auto-regressive)

Cross-attention

BART [Lewis+, 2019]
• Masked language model:

Encoder (bidirectional)

\[ \ldots \]

Decoder (Auto-regressive)

\[ \ldots \]

Cross-attention

Content plan (template)

BART [Lewis+, 2019]
• Masked language model:

Encoder (bidirectional)

Content plan (template)

Decoder (Auto-regressive)

Realization and Refinement

$[BOS] \ x_1 \ x_2 \ x_3 \ x_4 \ ... \ x_t \ [EOS]$
Proposal

• Prompt: Donald Trump is a communist.

• Content Plan:
  S1: A communist ▶ begin with ▶ coherent ideology ▶ [SEN]
  S2: [SEN]
  S3: No evidence ▶ any coherent ▶ held beliefs ▶ any topic ▶ [SEN]

• Template:
  S1: ______ a communist ______ begin with ________ coherent ideology ______
  S2: ________
  S3: ______ no evidence ________ any coherent ______ held beliefs ______ any topic ______
Proposal

• Template:
  S1: ___ a communist ___ begin with ___ coherent ideology ___
  S2: ___
  S3: ___ no evidence ___ any coherent ___ held beliefs ___ any topic ___

• Draft: (mask-and-fill)
  S1: well call him a communist you must begin with that Donald Trump has coherent ideology to begin with.
  S2: Which is unlikely.
  S3: There is no evidence to suggest Donald Trump has any coherent or commonly held beliefs on any topic.
Proposal

• Draft: (mask-and-fill)
  S1: well call him a communist you must begin with that Donald Trump has coherent ideology to begin with.
  S2: Which is unlikely.
  S3: There is no evidence to suggest Donald Trump has any coherent or commonly held beliefs on any topic.

• Refined: (Iterative Refinement)
  S1: To call him a communist you must begin with that he has some kind of coherent ideology in the first place.
  S2: He does not.
  S3: There is no evidence whatsoever that Trump has any coherent or commonly held beliefs on any topic.
Roadmap

• Motivation
• Related Work
• PAIR model
• Experiments
• Results
• Conclusion
Related Work

- **Text Planning in Neural Text Generation**
  - Plan as sequences or trees [Moryossef+, 2019; Puduppully+, 2019; Hua & Wang, 2019]

- **Controlled Text Generation**
  - Topic, style, entities [Wang+, 2017; See+, 2019; Fan+, 2018]
  - Syntax and semantics [Dusek & Jurcicek, 2016; Wen+, 2015]

- **Iterative Refinement**
  - Non-autoregressive machine translation [Lee+, 2018; Freitag+, 2019; Mansimov+, 2019; Kasai+, 2020]
  - Masked language model [Lawrence+, 2019; Ghazvininejad+, 2019]
Related Work

• **Text Planning in Neural Text Generation**
  • Plan as sequences or trees [Moryossef+, 2019; Puduppully+, 2019; Hua & Wang, 2019]

• **Controlled Text Generation**
  • Topic, style, entities [Wang+, 2017; See+, 2019; Fan+, 2018]
  • Syntax and semantics [Dusek & Jurcicek, 2016; Wen+, 2015]

• **Iterative Refinement**
  • Non-autoregressive machine translation [Lee+, 2018; Freitag+, 2019; Mansimov+, 2019; Kasai+, 2020]
  • Masked language model [Lawrence+, 2019; Ghazvininejad+, 2019]
Related Work

• **Text Planning in Neural Text Generation**
  • Plan as sequences or trees [Moryossef+, 2019; Puduppully+, 2019; Hua & Wang, 2019]

• **Controlled Text Generation**
  • Topic, style, entities [Wang+, 2017; See+, 2019; Fan+, 2018]
  • Syntax and semantics [Dusek & Jurcicek, 2016; Wen+, 2015]

• **Iterative Refinement**
  • Non-autoregressive machine translation [Lee+, 2018; Freitag+, 2019; Mansimov+, 2019; Kasai+, 2020]
  • Masked language model [Lawrence+, 2019; Ghazvininejad+, 2019]
Roadmap

• Motivation
• Related Work
• PAIR model
• Experiments
• Results
• Conclusion
PAIR Model

Input prompt $x$

Unordered keyphrase set $m$

Prompt: CMV. Donald Trump is a communist. Keyphrase set: {any coherent; any topic; begin with; coherent ideology; held beliefs ...}
Prompt: CMV. Donald Trump is a communist.
Keyphrase set: \{any coherent; any topic; begin with; coherent ideology; held beliefs \ldots\}

Text Plan: (a communist)\textsubscript{3} (begin with)\textsubscript{8} (coherent ideology)\textsubscript{15} \ldots
**Prompt:** CMV. Donald Trump is a communist.

**Keyphrase set:** \{any coherent; any topic; begin with; coherent ideology; held beliefs ...\}

**Text Plan:** (a communist)\textsubscript{3} (begin with)\textsubscript{8} (coherent ideology)\textsubscript{15} ...

**Template:** __ __ __ a communist __ __ begin with __ __ ...

---

**PAIR Model**

- **Input prompt** $x$
- **Unordered keyphrase set** $m$
- **Ordered keyphrase set** $m'$
- **Keyphrase token positions** $s$
PAIR Model

Input prompt $x$

Unordered keyphrase set $m$

Ordered keyphrase set $m'$

Keyphrase token positions $s$

Planner

Template $t^{(0)}$

Draft $y^{(1)}$

Generator
PAIR Model

Input prompt $x$

Unordered keyphrase set $m$

Planner

Ordered keyphrase set $m'$

Keyphrase token positions $s$

Generator

Template $t^{(0)}$

Draft $y^{(1)}$

Template $t^{(1)}$
PAIR Model

Input prompt $x$

Unordered keyphrase set $m$

Ordered keyphrase set $m'$

Keyphrase token positions $s$

Planner

Template $t^{(0)}$

Draft $y^{(1)}$

Template $t^{(1)}$

Generator
PAIR Model

Input prompt $x$

Unordered keyphrase set $m$

Ordered keyphrase set $m'$

Keyphrase token positions $s$

Planner

Generator

Template $t^{(r-1)}$

Draft $y^{(r)}$

Template $t^{(r)}$
Content Planning

**Input prompt**: \( x \)

**Keyphrase set**: \( m \)

Segment type 1

Segment type 2

[BOK] \( w_1 \) \( w_2 \) \( w_3 \)
Content Planning

Newly added output layer for position prediction

Input prompt $x$

Keyphrase set $m$

Segment type 1

Segment type 2

[Image of BERT model with segments and input prompt]
Content Planning

Segment embeddings to distinguish input and output

Input prompt $x$

Keyphrase set $m$

Segment type 1

Segment type 2

$w_1$ $w_2$ $w_3$ $w_4$

$s_1$ $s_2$ $s_3$ $s_4$

[BOK] $w_1$ $w_2$ $w_3$

BERT
Content Planning

Bi-directional self attention among input tokens

Input prompt $x$

Keyphrase set $m$

Segment type 1

Segment type 2

$[BOK] w_1 w_2 w_3 w_4$

$s_1 s_2 s_3 s_4$

Bi-directional self-attention among input tokens
Causal attention among output tokens

Content Planning

Input prompt $x$

Keyphrase set $m$

Segment type 1

Segment type 2

$BERT$

$w_1$, $w_2$, $w_3$, $w_4$

$s_1$, $s_2$, $s_3$, $s_4$

[BOK] $w_1$, $w_2$, $w_3$
• Initial template construction

Text Plan: (a communist)₃ (begin with)₈ (coherent ideology)₁₅ ...
Initial template construction

Text Plan: (a communist)$_3$ (begin with)$_8$ (coherent ideology)$_{15}$ ...

Template $t^{(0)}$: ___ ___ ___ a communist ___ ___ begin with ___ ___ ___ ...

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Initial template construction

Text Plan: (a communist)$_3$ (begin with)$_8$ (coherent ideology)$_{15}$ ...

Template $t^{(0)}$: __ __ __ a communist __ __ __ begin with __ __ __ ...

[MASK]
• BART decoding:

Encoder

Decoder

Input prompt $x$

Ordered keyphrases $m'$

Template $t^{(0)}$
• BART decoding:

```
Encoder
Input prompt x
Ordered keyphrases m'
Template t^{(0)}
```

```
Decoder
Draft y^{(1)}
```
Iterative Refinement

• Single-pass output tends to have low quality, due to the noises in planning and decoding.
Iterative Refinement

- Single-pass output tends to have low quality, due to the noises in planning and decoding.

- BART is naturally a denoising model!
• Masking tokens with low confidence [Ghazvininejad+, 2019]:

**Draft $y^{(1)}$**

S1: well call him a communist you must begin with that Donald Trump has coherent ideology to begin with.
S2: Which is unlikely.
S3: There is no evidence to suggest Donald Trump has any coherent or commonly held beliefs on any topic.
Iterative Refinement

• Masking tokens with low confidence [Ghazvininejad+, 2019]:

Draft $y^{(1)}$

S1: [mask] call him a communist you must begin with that [mask] [mask] has coherent ideology [mask] [mask] [mask].
S2: [mask] is unlikely.
S3: There is no evidence [mask] [mask] Donald Trump has any coherent or commonly held beliefs on any topic.
Iterative Refinement

- Masking tokens with low confidence [Ghazvininejad+, 2019]:

  Draft $y^{(2)}$
  
  S1: *To call him a communist you must begin with that he has coherent ideology in the first place.*
  
  S2: *That is unlikely.*
  
  S3: There is no evidence whatsoever that Donald Trump has any coherent or commonly held beliefs on any topic.
Iterative Refinement

• Masking tokens with low confidence (Ghazvininejad+, 2019):

Decoding: nucleus sampling for three times, pick the one with the lowest perplexity (by a GPT-2).

Well call him a communist, you must begin with that Donald Trump...
• Masking tokens with low confidence [Ghazvininejad+, 2019]:

Mask k tokens, where k linearly decrease by iterations.
Iterative Refinement

• Masking tokens with low confidence [Ghazvininejad+, 2019]:

If to call him a communist you need to begin with the idea that...
Iterative Refinement

- Masking tokens with low confidence [Ghazvininejad+, 2019]:

BART is capable to delete and insert extra tokens, thus refinement is flexible.
Iterative Refinement

• Masking tokens with low confidence [Ghazvininejad+, 2019]:

We enforce keyphrases to be mentioned at or nearby the predicted positions.
BART decoding:

- Input prompt $x$
- Ordered keyphrases $m'$
- Template $t^{(r-1)}$
- Draft $y^{(r)}$

Mask tokens
Roadmap

• Motivation
• Related Work
• PAIR model
• Experiments
• Results
• Conclusion
Experiments

• Tasks:
  • Argument generation (ARGGEN)
  • Opinion article generation (OPINION)
  • News report generation (NEWS)
Experiments

• Tasks:
  • Argument generation (ARGGEN)

CMV: Donald Trump is a communist

Delta(s) from OP

Change My View (CMV)
r/changemyview

His international friends consists of the two most notorious communists in the modern world. Vladimir Putin member of the Communist Party of the Soviet Union (1975–91) and Xi Jinping emperor of Chine and General Secretary of the Communist Party of China. D.T. (Donald Trump) constantly praises China and regards Putin as a friend (Probably the terme friend is far fetched, but the fact that D.T. asked for russia to be reintegrated on the G-7 and also his one on one meetings with Putin are proof of friendship with the communist kaiser).
Experiments

• Tasks:
  • Argument generation (**ARGGEN**)

CMV: Donald Trump is a communist

His international friends consists of the two most notorious communists in the modern world. Vladimir Putin member of the Communist Party of the Soviet Union (1975–91) and Xi Jinping emperor of Chine and General Secretary of the Communist Party of China. D.T. (Donald Trump) constantly praises China and regards Putin as a friend (Probably the terme friend is far fetched, but the fact that D.T. asked for russia to be reintegrated on the G-7 and friendship with the communist kaiser).

Saying that he is a communist presumes to begin with that Donald Trump has some kind of coherent ideology in the first place. He doesn't. There is no evidence that he Trump has any coherent or firmly held beliefs on any topic.

White-Afro-American 🖇 1 point · 2 years ago · edited 2 years ago

You actually changed my view to Donald Trump acts like a communist.

Thank you for your answer. Δ
Experiments

• Tasks:
  • Argument generation (ARGGEN)
  • Opinion article generation (OPINION)
  • News report generation (NEWS)

CMV: Donald Trump is a communist

OP -> input prompt

High quality replies -> Keyphrases, target

Saying that he is a communist presumes to begin with that Donald Trump has some kind of coherent ideology in the first place. He doesn't. There is no evidence that he Trump has any coherent or firmly held beliefs on any topic.

You actually changed my view to Donald Trump acts like a communist.

Thank you for your answer. Δ
### Experiments

**Tasks:**

- Argument generation (**ARGGEN**)
  - Opinion article generation (**OPINION**)
  - News report generation (**NEWS**)

<table>
<thead>
<tr>
<th>Size</th>
<th>Prompt Len.</th>
<th># Keyphrases</th>
<th>Target Len.</th>
</tr>
</thead>
<tbody>
<tr>
<td>56,504</td>
<td>19.4</td>
<td>20.6</td>
<td>116.6</td>
</tr>
</tbody>
</table>

Saying that he is a communist presumes to begin with that Donald Trump has some kind of coherent ideology in the first place. He doesn’t. There is no evidence that he Trump has any coherent or firmly held beliefs on any topic.
Experiments

• Tasks:
  • Argument generation (ARGGEN)
  • Opinion article generation (OPINION)
  • News report generation (NEWS)

Articles from the New York Times corpus [Sandhaus, 2008].

“Top/Opinion” -> OPINION
“Top/News” -> NEWS
Experiments

• Tasks:
  • Argument generation (ARGSGEN)
  • Opinion article generation (OPINION)
  • News report generation (NEWS)

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Prompt Len.</th>
<th># Keyphrases</th>
<th>Target Len.</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPINION</td>
<td>104,610</td>
<td>6.1</td>
<td>19.0</td>
<td>205.6</td>
</tr>
<tr>
<td>NEWS</td>
<td>239,959</td>
<td>7.0</td>
<td>30.3</td>
<td>282.7</td>
</tr>
</tbody>
</table>
Experiments

• Baselines:
  • $\text{SEQ2SEQ}$: directly generates target from the prompt
  • $\text{KPSSEQ2SEQ}$: generates target from prompt and unordered keyphrase set
Experiments

• Baselines:
  • $\text{SEQ2SEQ}$: directly generates target from the prompt
  • $\text{KPSSEQ2SEQ}$: generates target from prompt and unordered keyphrase set
Experiments

• Comparisons:
  • $\text{PAIR}_{\text{full}}$: our proposed model with iterative refinement

![Diagram](Diagram.png)
Experiments

Comparisons:
- $\text{PAIR}_{\text{full}}$: our proposed model with iterative refinement
- $\text{PAIR}_{\text{full}}$ without refinement: one-pass generation

Diagram:
- Encoder
  - Input prompt $x$
  - Keyphrase list $m'$
  - Initial template
- Decoder
  - Draft $y^{(1)}$
Experiments

- **Comparisons:**
  - $\text{PAIR}_{\text{full}}$: our proposed model with iterative refinement
  - $\text{PAIR}_{\text{full}}$ without refinement: one-pass generation
  - $\text{PAIR}_{\text{light}}$: replace initial template with all masks

```
Encoder
Input prompt $x$
Keyphrase list $m'$
[mask] * N

Decoder
Draft $y^{(r)}$
Model without keyphrase position information.
```
Experiments

• Comparisons:
  • $\text{PAIR}_{\text{full}}$: our proposed model with iterative refinement
  • $\text{PAIR}_{\text{full}}$ without refinement: one-pass generation
  • $\text{PAIR}_{\text{light}}$: replace initial template with all masks
  • $\text{PAIR}_{\text{light}}$ without refinement
Roadmap

• Motivation
• Related Work
• PAIR model
• Experiments
• Results
• Conclusion
Results

- Automatic Evaluation – ARGGEN

<table>
<thead>
<tr>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>METEOR</th>
<th>Length</th>
</tr>
</thead>
</table>

Results with ground-truth content plans.
## Results

- **Automatic Evaluation – ARGGen**

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>METEOR</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQ2SEQ</td>
<td>0.76</td>
<td>13.80</td>
<td>9.36</td>
<td>97</td>
</tr>
<tr>
<td>KPSEQ2SEQ</td>
<td>6.78</td>
<td>19.43</td>
<td>15.98</td>
<td>97</td>
</tr>
</tbody>
</table>
## Results

- **Automatic Evaluation – ARGGEN**

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>METEOR</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQ2SEQ</td>
<td>0.76</td>
<td>13.80</td>
<td>9.36</td>
<td>97</td>
</tr>
<tr>
<td>KPSEQ2SEQ</td>
<td>6.78</td>
<td>19.43</td>
<td>15.98</td>
<td>97</td>
</tr>
<tr>
<td><strong>PAIR_{light}</strong></td>
<td>26.38</td>
<td>47.97</td>
<td>31.64</td>
<td>119</td>
</tr>
<tr>
<td><strong>PAIR_{light} w/o refine</strong></td>
<td>25.17</td>
<td>46.84</td>
<td>31.31</td>
<td>120</td>
</tr>
<tr>
<td><strong>PAIR_{full}</strong></td>
<td>36.09</td>
<td>56.86</td>
<td>33.30</td>
<td>102</td>
</tr>
<tr>
<td><strong>PAIR_{full} w/o refine</strong></td>
<td>34.09</td>
<td>55.42</td>
<td>32.74</td>
<td>101</td>
</tr>
</tbody>
</table>
## Results

- **Automatic Evaluation – ARGGEN**

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>METEOR</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQ2SEQ</td>
<td>0.76</td>
<td>13.80</td>
<td>9.36</td>
<td>97</td>
</tr>
<tr>
<td>KPSEQ2SEQ</td>
<td>6.78</td>
<td>19.43</td>
<td>15.98</td>
<td>97</td>
</tr>
<tr>
<td>PAIR&lt;sub&gt;light&lt;/sub&gt;</td>
<td>26.38</td>
<td>47.97</td>
<td>31.64</td>
<td>119</td>
</tr>
<tr>
<td>PAIR&lt;sub&gt;light&lt;/sub&gt; &lt;sub&gt; w/o refine&lt;/sub&gt;</td>
<td>25.17</td>
<td>46.84</td>
<td>31.31</td>
<td>120</td>
</tr>
<tr>
<td>PAIR&lt;sub&gt;full&lt;/sub&gt;</td>
<td><strong>36.09</strong></td>
<td><strong>56.86</strong></td>
<td><strong>33.30</strong></td>
<td>102</td>
</tr>
<tr>
<td>PAIR&lt;sub&gt;full&lt;/sub&gt; &lt;sub&gt; w/o refine&lt;/sub&gt;</td>
<td>34.09</td>
<td>55.42</td>
<td>32.74</td>
<td>101</td>
</tr>
</tbody>
</table>

Position information helps (full model).
## Results

- **Automatic Evaluation – ARGGEN**

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>METEOR</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQ2SEQ</td>
<td>0.76</td>
<td>13.80</td>
<td>9.36</td>
<td>97</td>
</tr>
<tr>
<td>KPSEQ2SEQ</td>
<td>6.78</td>
<td>19.43</td>
<td>15.98</td>
<td>97</td>
</tr>
<tr>
<td>PAIR\textsubscript{light}</td>
<td>26.38↑</td>
<td>47.97↑</td>
<td>31.64↑</td>
<td>119</td>
</tr>
<tr>
<td>PAIR\textsubscript{light} w/o refine</td>
<td>25.17</td>
<td>46.84</td>
<td>31.31</td>
<td>120</td>
</tr>
<tr>
<td>PAIR\textsubscript{full}</td>
<td>36.09↑</td>
<td>56.86↑</td>
<td>33.30↑</td>
<td>102</td>
</tr>
<tr>
<td>PAIR\textsubscript{full} w/o refine</td>
<td>34.09</td>
<td>55.42</td>
<td>32.74</td>
<td>101</td>
</tr>
</tbody>
</table>

Refinement improves all metrics.
## Results

### Automatic Evaluation – **OPINION**

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>METEOR</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQ2SEQ</td>
<td>1.42</td>
<td>15.97</td>
<td>10.97</td>
<td>156</td>
</tr>
<tr>
<td>KPSEQ2SEQ</td>
<td>11.38</td>
<td>22.75</td>
<td>18.38</td>
<td>164</td>
</tr>
<tr>
<td>$\text{PAIR}_{\text{light}}$</td>
<td>16.27↑</td>
<td>33.30↑</td>
<td>24.32↑</td>
<td>210</td>
</tr>
<tr>
<td>$\text{PAIR}_{\text{light}}$ w/o refine</td>
<td>15.45</td>
<td>32.35</td>
<td>24.11</td>
<td>214</td>
</tr>
<tr>
<td>$\text{PAIR}_{\text{full}}$</td>
<td>23.12↑</td>
<td>40.53↑</td>
<td>24.73↑</td>
<td>167</td>
</tr>
<tr>
<td>$\text{PAIR}_{\text{full}}$ w/o refine</td>
<td>22.17</td>
<td>39.71</td>
<td>24.65</td>
<td>169</td>
</tr>
</tbody>
</table>
### Results

**Automatic Evaluation – NEWS**

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>METEOR</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SEQ2SEQ</strong></td>
<td>1.11</td>
<td>15.60</td>
<td>10.10</td>
<td>242</td>
</tr>
<tr>
<td><strong>KPSEQ2SEQ</strong></td>
<td>11.61</td>
<td>21.05</td>
<td>18.61</td>
<td>286</td>
</tr>
<tr>
<td><strong>PAIR\textsubscript{light}</strong></td>
<td>28.03↑</td>
<td>43.39↑</td>
<td>27.70↑</td>
<td>272</td>
</tr>
<tr>
<td><strong>PAIR\textsubscript{light} w/o refine</strong></td>
<td>27.32</td>
<td>43.08</td>
<td>27.35</td>
<td>278</td>
</tr>
<tr>
<td><strong>PAIR\textsubscript{full}</strong></td>
<td>34.37↑</td>
<td>51.10↑</td>
<td>29.50↑</td>
<td>259</td>
</tr>
<tr>
<td><strong>PAIR\textsubscript{full} w/o refine</strong></td>
<td>33.48</td>
<td>50.27</td>
<td>29.26</td>
<td>260</td>
</tr>
</tbody>
</table>
Results

• Change across iterations:

- BLEU-4
- ROUGE-L
- Perplexity

![Graphs showing changes across iterations for BLEU-4, ROUGE-L, and Perplexity for different datasets: ARGGen, Opinion, News.](image-url)
Results

- Results with system predicted content plans:

![Graphs showing BLEU-4 and METEOR scores for ArgGen, Opinion, and News categories. The graphs compare results from KPSeq2seq, PAIRlight w/o refine, PAIRfull w/o refine, and PAIRfull.]
Results

• Human Evaluation
  • **Fluency** (1-5): if the output is grammatical and readable
  • **Coherence** (1-5): if the information organization is natural and logical
  • **Relevance** (1-5): if the topic is related to the prompt and whether the stance is correct
Results

- **Human Evaluation**
  - **Fluency** (1-5): if the output is grammatical and readable
  - **Coherence** (1-5): if the information organization is natural and logical
  - **Relevance** (1-5): if the topic is related to the prompt and whether the stance is correct

<table>
<thead>
<tr>
<th></th>
<th>Fluency</th>
<th>Coherence</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPSEQ2SEQ</td>
<td>4.63</td>
<td>3.28</td>
<td>2.79</td>
</tr>
<tr>
<td>PAIR_{light}</td>
<td>4.75</td>
<td>3.97</td>
<td>3.85</td>
</tr>
<tr>
<td>PAIR_{full}</td>
<td>4.46</td>
<td>3.76</td>
<td>3.79</td>
</tr>
</tbody>
</table>
Prompt: CMV. The EU needs its own independent armed forces

I don’t see *the primary benefit* of *a military* under *the authority* of *EU institutions*. Instead, I think *the EU* should set up its own armed forces with *UN Security Council authorisation*, to *the EU member states*, or *any subset* of EU member states that wishes not to participate. Ultimately, *the primary advantage* of *a EU military* is that *military operations conducted under the authority* of EU institutions already have *a level of legitimacy* comparable to *individual national states*, [...]
Prompt: The World’s Best Job

I was happy to see that travel writers are getting better. Many of those travel writers and research guides in Goa and Romania had a rough time there, but the vast majority of them, as is all travel writers. They all work hard and do a lot of research trips all over the world. They research trips, are hectic, and they are much more than the ideal job or hobby. […]
Kraft Foods expanded its agency roster by naming McGarry Bowen in New York to create campaigns for salad dressings, mayonnaise and barbecue sauces bearing the Kraft brand name. Spending was estimated at $10 million. The decision to expand its relationship with McGarry Bowen is part of a trend in which big marketers like big marketers like Kraft shift creative assignments from small, independent agencies to larger agencies at the same time. […]
What is updated during refinement?

• About 85% refinement operations changed content, 1~5% added tokens, 2~6% deleted tokens.
What is updated during refinement?

• About 85% refinement operations changed content, 1~5% added tokens, 2~6% deleted tokens.

• 38%~41% masks are put on unigrams, 22% on bigrams, 13~15% on trigrams.
What is updated during refinement?

- **Top frequent changes:**

<table>
<thead>
<tr>
<th>ARGGen</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>because</td>
<td>,</td>
<td>Feb</td>
</tr>
<tr>
<td>It</td>
<td>That</td>
<td>Feb</td>
</tr>
<tr>
<td>?</td>
<td>.</td>
<td>is</td>
</tr>
<tr>
<td>their</td>
<td>your</td>
<td>does</td>
</tr>
<tr>
<td>need</td>
<td>have</td>
<td>from</td>
</tr>
<tr>
<td>actually</td>
<td>really</td>
<td>more</td>
</tr>
<tr>
<td>may</td>
<td>would</td>
<td>can</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OPINION</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb</td>
<td>Jan</td>
<td></td>
</tr>
<tr>
<td>Dec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>was</td>
<td></td>
<td>from</td>
</tr>
<tr>
<td>is</td>
<td>did</td>
<td></td>
</tr>
<tr>
<td>from</td>
<td>announced</td>
<td></td>
</tr>
<tr>
<td>more</td>
<td>less</td>
<td></td>
</tr>
<tr>
<td>can</td>
<td>cannot</td>
<td></td>
</tr>
<tr>
<td>first</td>
<td>second</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NEWS</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>added</td>
<td>said</td>
<td></td>
</tr>
<tr>
<td>with</td>
<td>and</td>
<td></td>
</tr>
<tr>
<td>from</td>
<td>to</td>
<td></td>
</tr>
<tr>
<td>did</td>
<td>would</td>
<td></td>
</tr>
<tr>
<td>announced</td>
<td>said</td>
<td></td>
</tr>
<tr>
<td>While</td>
<td>Although</td>
<td></td>
</tr>
</tbody>
</table>
What is updated during refinement?

• Top frequent changes:

<table>
<thead>
<tr>
<th>ARGGEN</th>
<th>OPINION</th>
<th>NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>New</td>
<td>Old</td>
</tr>
<tr>
<td>because</td>
<td>,</td>
<td>Feb</td>
</tr>
<tr>
<td>It</td>
<td>That</td>
<td>Feb</td>
</tr>
<tr>
<td>?</td>
<td>.</td>
<td>is</td>
</tr>
<tr>
<td>their</td>
<td>your</td>
<td>does</td>
</tr>
<tr>
<td>need</td>
<td>have</td>
<td>from</td>
</tr>
<tr>
<td>actually</td>
<td>really</td>
<td>more</td>
</tr>
<tr>
<td>may</td>
<td>would</td>
<td>can</td>
</tr>
</tbody>
</table>

Refinement can produce factual errors!
Is PAIR better at capturing discourse?

• Study on RST tree depth using DPLP [Ji & Eisenstein, 2014]
Is PAIR better at capturing discourse?

• Study on RST tree depth using DPLP [Ji & Eisenstein, 2014]
Is PAIR better at capturing discourse?

- Discourse marker accuracy
Is PAIR better at capturing discourse?

- Discourse marker accuracy

PAIR is better at Contingency than Comparison and Expansion.
Conclusion

• We propose a content-controlled generation framework with separate planning and realization.

• Our model leverages the masked language model for iterative refinement to improve the output quality.

• Both automatic and human evaluation show the effectiveness of our model over three tasks.
Thanks!

https://xinyuhua.github.io/Resources/emnlp20/

hua.x@northeastern.edu

https://github.com/XinyuHua/pair-emnlp2020