Plan ahead: Self-Supervised Text Planning for Paragraph Completion Task

Dongyeop Kang*
dongyeopk@berkeley.edu
UC Berkeley

Eduard Hovy
hovy@cs.cmu.edu
Carnegie Mellon University

Abstract

Despite the recent success of contextualized language models on various NLP tasks, language model itself cannot capture textual coherence of a long, multi-sentence document (e.g., a paragraph). Humans often make structural decisions on what and how to say about before making utterances. Guiding surface realization with such high-level decisions and structuring text in a coherent way is essentially called a planning process. Where can the model learn such high-level coherence? A paragraph itself contains various forms of inductive coherence signals called self-supervision in this work, such as sentence orders, topical keywords, rhetorical structures, and so on. Motivated by that, this work proposes a new paragraph completion task PAR-COM; predicting masked sentences in a paragraph. However, the task suffers from predicting and selecting appropriate topical content with respect to the given context. To address that, we propose a self-supervised text planner SSPPlanner that predicts what to say first (content prediction), then guides the pretrained language model (surface realization) using the predicted content. SSPPlanner outperforms the baseline generation models on the paragraph completion task in both automatic and human evaluation. We also find that a combination of noun and verb types of keywords is the most effective for content selection. As more number of content keywords are provided, overall generation quality also increases.

1 Introduction

One may think textual coherence can be achieved from a gigantic language model trained on massive data. This might be true in simple cases, such as generating short replies (Kannan et al., 2016), but not in a long, multi-sentence generation. This is mainly because per-word predictions from the autoregressive models can not capture the long-term flow of text, while humans often make structural decisions on what and how to say about before they speak (Byrne, 1979; McKeown, 1985; Hovy, 1990; Swan, 2002; Kang, 2020). Guiding the surface-level realization with such high-level decisions and coherently structuring output text is called a planning process.

Where can the model learn such high-level decisions related to long-term coherence? A written paragraph itself can be a pot of golden resources, containing various forms of inductive coherence signals. Different types of coherence signals in a paragraph have been studied and used in many different ways: a sequence of words or sentences (Doblin et al., 2019; Radford et al., 2019), a discourse structure of a text (Appelt, 1982; Hovy, 1991; Kang et al., 2019), an order of sentences (Chambers and Jurafsky, 2008; Barzilay and Lapata, 2008), topic introduction, co-reference, a sequence of events (Tomkins, 1978; Schank and Abelson, 2013), and more. In this work, we primarily focus on the effect of topical content in text planning.

Despite the recent advances of contextualized language models (Devlin et al., 2019; Radford et al., 2019), the lack of appropriate tasks makes it difficult to evaluate generation models’ long-term coherence. Prior tasks fall into classification or ranking problems, such as narrative close task (Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016), sentence ordering (Barzilay and Lapata, 2008), and next sentence prediction (Devlin et al., 2019). Some recent works focused on designing generation tasks: story generation (Fan et al., 2019), text infilling (Huang et al., 2019; Fedus et al., 2018; Hua and Wang, 2019), or paragraph bridging (Kang et al., 2019). However, most of them suffer from predicting appropriate topical content given limited context, due to the limited usage of self-supervision.

*This work was done while DK was at CMU.
signals from the paragraph.

This work proposes a new open-ended paragraph completion task; PARCOM; predicting the masked sentences in a paragraph. Unlike the prior works, our task uses two effective ways of self-supervision learnt from a written paragraph itself: (1) we augment more training instances via permutation masking and (2) resolve the context sparsity problem by providing a set of ground-truth content keywords and predicting them directly from context at testing time.

For the task, we propose a self-supervised text planner (SSPlanner) that explicitly predicts content keywords (content prediction) from context and guides the pretrained language model (surface-realization) using the predicted content. The distribution of predicted keywords is then combined with the distribution of words in the language model using copy mechanism (See et al., 2017). The predicted content keywords are an approximation of topical intents by the generator, providing a hint to guide the surface realizer to bridge the coherency gap between the given context and text to generate. Overall, SSPlanner combines two advantages; micro-level language fluency from the pre-trained language model (bottom-up) and macro-level content choice controlled by the macro-level planning (top-down). Our experiment shows that SSPlanner achieves significant improvements over the baselines in both automatic and human evaluation.

2 Related Work

We first categorize a wide range of long-term coherent generation tasks (Table 1), based on their inclusion relationship (C-T) between the given context (C) and target to predict (T).

<table>
<thead>
<tr>
<th>C-T</th>
<th>Tasks</th>
<th>Content Selection</th>
<th>Content Planning</th>
<th>Content Ordering</th>
<th>Surface Realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>Data-to-text</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>☒</td>
<td>Summarization</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>☒</td>
<td>Paraphrasing</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>☒</td>
<td>StoryGen, Text Infilling</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>☒</td>
<td>Bridging, PARCOM (ours)</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
</tbody>
</table>

Table 1: Comparison of generation tasks by different inclusion relationships between Context and Target.

C ⊆ T: Data-to-text produces text from structured data (e.g., table). Moryossef et al. (2019); Puduppully et al. (2019); Shen et al. (2019); Miculich et al. (2019) combine content planning with surface realization. However, since content is explicitly provided as a data form, the planner mostly orders and structures, not prediction.

C ⊇ T: In abstractive summarization, all context information is entirely given in the source document, as a superset of target summaries to predict. Thus, generation only pays attention to abstracting the context into a shorter form instead of content prediction or ordering.

C ≈ T: Paraphrasing is transforming surface patterns of text while preserving its semantics. Fu et al. (2019) used variational autoencoders for surface realization with a latent bag of words model for differentiable content planning, where content to generate itself is given in context, not requiring any content planning.

C ⊥⊥ T: Story generation (Fan et al., 2019), text infilling (Fedus et al., 2018; Huang et al., 2019), paragraph bridging (Kang et al., 2019), and our proposed PARCOM are very challenging tasks where context and target have no overlap (open-ended), but they should be coherently connected. Table 2 categorize various generation models applied on the open-ended tasks (C⊥⊥T), based on its self-supervision types:

<table>
<thead>
<tr>
<th>Models (⊂⊥⊥)</th>
<th>Content Prediction</th>
<th>Content Guidance</th>
<th>Content Masking</th>
<th>Bidirect. Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keskar et al. (2019)</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Fan et al. (2019)</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Huang et al. (2019)</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Hua and Wang (2019)</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Kang et al. (2019)</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>SSPlanner (ours)</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
</tbody>
</table>

Table 2: Comparison of generation models in C⊥⊥T tasks by different self-supervision types.

Keskar et al. (2019) conditioned language models with topical words to control the target text. Fan et al. (2019) developed a surface realizer on anonymized entities using semantic role labeling. Hua and Wang (2019) used pre-extracted topics to guide a generator to produce stylized argumentation text. However, they are given the topical content as input (content guidance), while our SSPlanner directly predicts plan words from context (content prediction).

Fedus et al. (2018); Huang et al. (2019) developed various methods for text infilling task. Very similar to our task, Kang et al. (2019) developed language models informed by discourse relations on the bridging task; given the first and last sentences, predicting the intermediate sentences (bidirectional flow). However, they did not explicitly
predict content words given context nor use them as a self-supervision signal in training. Unlike random masking in Keskar et al. (2019); Huang et al. (2019), we propose a better data augmentation training method via permutation masking.

3 PARCOM: Paragraph Completion Task from Self-Supervision Signals

Our task is motivated by the recently proposed task; paragraph bridging (Kang et al., 2019), predicting intermediate sentences of a paragraph, given the first and the last sentences. To prevent generation becoming too divergent from the context in story or prompt generation (Fan et al., 2019), the bridging task restricts generation to end with the last sentence given, provided as an ending goal for generation.

However, in the bridging task, the objective is to generate text by coherently linking the two extreme sentences, making the task itself too challenging even for human\(^1\). For instance, the first and last sentences are too sparse to generate multiple (from 2 to 5) target sentences, increasing divergence of generation exponentially. Also, data usage in (Kang et al., 2019) is very inefficient; training a single instance per paragraph.

To address those issues, we propose a new paragraph completion task PARCOM by maximizing self-supervision presented in a paragraph itself (Figure 1). We describe two types of self-supervisions: (1) masking a fixed-length of consecutive sentences in any position over a paragraph to maximize usage of a paragraph and (2) extracting partial keywords of the masked text as plan keywords to resolve the content sparsity problem. Mainly, we learn the patterns between the context and the plan keywords in training and at testing time predict the plan keywords, and guide the surface generator (§4).

3.1 Data Augmentation via Permutation Masking

Our work is motivated by word masking, in training contextualized language models (Devlin et al., 2019), but extending it to sentence-level for learning longer coherence.

Let \( t \) be the number of targets, masked sentences to predict and \( c \) be the number of unmasked, context sentences given, where \( l = t + c \) is the total length of a paragraph. For instance, in Figure 1, we have a \( l=5 \) length paragraph. We restrict the number of context sentences to be larger than the number of target sentences \( (c > t) \), to avoid context become too sparse. Also, we produce a total of \( 5+4=9 \) training instances, making use of data more efficient.

3.2 Denser Context by Plan Extraction

We provide extra partial information as a set of keywords to guide the surface generator. This is motivated by data-to-text tasks, but our plans are topical content instead of structured data.

We then question what types of plan keywords are the most effective for completing the paragraph. We extract keywords using various keyword extraction systems:

- **Off-the-shelf** systems extract keywords for each sentence using the three off-the-shelf systems: YAKE (Campos et al., 2020) using statistical features (e.g., TF, IDF), RAKE (Rose et al., 2010) using graph-based features (e.g., word degree), and PositionRank (Florescu and Caragea, 2017) using position-based PageRank. Then we choose duplicate keywords by majority voting.

- **Syntactic** features (e.g., part-of-speech tags, named entities (Fan et al., 2019), events
(Tomkins, 1978)) are often regarded as the most salient topical content in generation. Using off-the-shelf Part-of-speech (PoS) tagger\(^2\), we extract three types of syntactic features: *nouns*, *verbs*, and *nouns+verbs*.

- **Attention** weights are used to capture context-aware keywords. We use the pre-trained BERT (Devlin et al., 2019) to encode context and target text, then average the attention weights of context words with respect to each target word. We only use the first head’s attentions, then average them over all 12 layers\(^3\). We finally choose words with the maximum weight except for the special tokens (e.g., [CLS]) and punctuation marks.

We set the maximum number of keywords per sentence (*nkps*) to 5. Some extractors output an empty keyword list, so the number of keywords across the systems is different. Our keywords are always unigrams. In case they are not unigrams, we split them by whitespaces and use individual unigrams as unique keywords. If the target text has multiple sentences, we combine all keywords from the sentences and randomly shuffle them. The plan keywords extracted are only provided while training our plan predictor, but not at test time. At testing time, we explicitly predict the keywords given context.

### 4 Self-supervised Text Planning (SSPlanner)

SSPlanner has various self-supervision modules that learn coherence signals from a paragraph itself: *surface realizer* (language model) by learning from a sequence of words, *next sentence predictor* by learning from a sequence of two consecutive sentences, *sentence position embeddings* by learning from an order of context sentences, *plan predictor* by learning from the relationship between the given context and important keywords used in the generation of the target text, and *content guidance* by learning from whether the predicted plan keywords are used or not in the target (See Figure 2).

Our planner is motivated by the two-stage generation framework (Moryossef et al., 2019; Miculicich et al., 2019; Fu et al., 2019; Hua and Wang, 2019). While in prior works, the content is explicitly given from the dataset or task itself, our plan predictor in SSPlanner predicts the plan keywords only from the given context, by learning the topical relationship between context and content in target from training data.

Given \( l \) length of a paragraph \( s_1,...,s_l \) where each sentence \( s \) consists of a \( n \) number of words \( s = w_1,...,w_n \), PARCOM splits it into the context sentences \( x = s_1,...,s_{j-1},s_{j+1},...,s_n \) and \( t \) target sentences to predict \( y = s_j,...,s_{j+t-1} \). For each target sentence, \( p \) number of plan keywords \( k_{1...p} \) for arbitrary target sentence \( s_j \) are given only at training time. The plan keywords are chosen from the entire vocabulary \( \forall W \) and later combined with word distribution from the language model. We describe each self-supervision module in SSPlanner as follows:

**Surface realization with pre-trained language models.** We use two different types of transformer-based language models: BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019). While GPT2 is trained on bidirectionally tied language modeling, BERT is trained on masked language modeling. For BERT, we use the sequential sampling method (Wang and Cho, 2019). Using them, we encode context \( x \) and output the hidden representation \( h_{j,i} = \mathbf{f}(h_{j-1,i}, x_{s_{j-1:(j,i)}}) \) for \( j \)th word in \( j \)th sentence, where \( \mathbf{f} \in \{ \text{BERT}, \text{GPT2} \} \) is the transformer language model. We then output the sentence vector \( h_i \) by averaging all word vectors in a sentence.

**Sentence position embedding.** We concatenate the encoded sentence representation with its sentence position embedding. By adding the sentence position embeddings into context encoding, the model is aware of where the context sentence came from (e.g., from the first or last). Compared to the simple concatenation of them (Kang et al., 2019), our sentence position embedding helps better learn the bi-directional coherence. The context vector’s final representation is then \( h^c = \frac{1}{n} \sum_i h_i; \text{pos}_i^n \) where \( n \) is the number of sentences in a text and \( \text{pos}_i^n \) is the position embedding of \( i \)th sentence in the context paragraph.

**Plan prediction.** This work assumes that high-level plan words consist of bag-of-words (Fu et al., 2019), so that the model directly predicts the plan keywords from the vocabulary used in surface realization. We calculate the plan probabilities over the entire vocabularies \( V \) given the context vector \( h^c \) and choose the \( p \) number of keywords with maximum probability estimates over vocabulary: \( \tilde{p}_{k \in \mathcal{V}} = \text{softmax}(h^c W^{cy}) \) where \( \mathcal{V} \) is the vocabu-

\(^2\)https://spacy.io/

\(^3\)Vig (2019) observed that which layers or heads are important for syntactic and semantic tasks.
lary from the training data and $W^c$ is the trainable model parameter. We do not control any explicit cut-off in the $p_{k \in \mathcal{V}}$ in order to make the distribution differentiable. The objective is then:

$$L_{\text{plan}} = - \sum_{k \in \mathcal{V}} \log p_k \log \hat{p}_k$$  \hspace{1cm} (1)

where the loss is calculated by cross-entropy, $\hat{p}$ is the estimated probability distribution over vocabulary and $p_k$ is the true one-hot distribution over plan keywords extracted from the extraction algorithms (i.e., $[0,1,0,1]$ over $\mathcal{V}$).

**Next sentence prediction.** Motivated by Devlin et al. (2019), we also add an auxiliary task of predicting whether the target sentence is related to context or not. For negative samples, PARCOM assigns 50% of random target sentences. We optimize $\hat{p}_{\text{next}} = \text{softmax}(W^c h^c)$ where $W^c$ is the trainable parameter for the binary classification. Next sentence prediction’s objective is then:

$$L_{\text{next}} = - \sum_{j} p_{\text{next}} \log \hat{p}_{\text{next}}$$  \hspace{1cm} (2)

where the loss is calculated by binary cross-entropy, $p_{\text{next}}$ is the true label for next sentences and $\hat{p}_{\text{next}}$ is the predicted label.

**Content guidance.** We combine two distributions between plan predictions and language modeling through copy mechanism following the pointer-generator (See et al., 2017). For $j^{th}$ sentence, we learn the probability of choosing the plan keyword or the word from language modeling based on context vectors, plan keyword distributions, and sentence position embedding of target sentences:

$$P_{\text{plan}}(v_k) = \sigma(W^k [h^c; \hat{p}_k; \text{pos}_j])$$, where $\sigma$ is a sigmoid function, $W^k$ is the trainable parameter, and $v \in [0,1]$ is a probability of whether choosing the plan keyword or not.

We then decode each target sentence using the same language model decoder: $s_j = g(s_{j-1}, \hat{y}_{j-1})$, where $g \in \{\text{BERT, GPT2}\}$ is the language model decoder and $s$ is its output hidden state. We can obtain the attention over plan keywords $k$:

$$\alpha_k^{\text{plan}} = \text{softmax}(\hat{p}_k W^{k|j} [s_j; \text{pos}_j])$$  \hspace{1cm} (3)

where $W^{k|j}$ is the trainable parameter. Lastly, we combine the distribution of plan probabilities $P_{\text{plan}}$ and word probabilities in decoding $P_{\text{lm}}$.

$$P(y) = P_{\text{plan}} \sum_k (\alpha_k^{\text{plan}}) + (1 - P_{\text{plan}})P_{\text{lm}}(y)$$  \hspace{1cm} (4)

The objective of the pointer-generator is then:

$$L_{\text{gen}} = - \sum_{i \in \mathcal{X}, j=1..n} P(\hat{y}_{i,j}) \log P(y_{i,j}^*)$$  \hspace{1cm} (5)

**Final objective.** The final objective of our training is to minimize the three objectives; plan prediction, next sentence prediction, and pointer generation, together:

$$L_{\text{SPP}} = \lambda_{\text{plan}} L_{\text{plan}} + \lambda_{\text{next}} L_{\text{next}} + L_{\text{gen}}$$  \hspace{1cm} (6)
where the weighting terms; $\lambda_{plan}$ and $\lambda_{next}$, are obtained through the cross-validation.

5 Experiment

We answer three questions in our experiments: Q1. Does SSPlanner help produce a more coherent generation in PARCOM? If so, which of the self-supervision modules are the most helpful? Q2. What types of plan keywords (e.g., noun, verb, attention) are most effective in terms of generation quality? How many keywords given are the most helpful? Q3. Is PARCOM a valid generation task to measure text coherence?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>#Sent.</th>
<th>#Para.</th>
<th>Length</th>
<th>#Inst.</th>
<th>#Keyw.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fantasy</td>
<td>book</td>
<td>1.6M</td>
<td>352K</td>
<td>4.7</td>
<td>21M</td>
<td>3.2-19M</td>
</tr>
<tr>
<td>Romance</td>
<td>book</td>
<td>5.3M</td>
<td>1.1M</td>
<td>4.6</td>
<td>67M</td>
<td>4.6-62M</td>
</tr>
<tr>
<td>WikiText</td>
<td>wiki</td>
<td>510K</td>
<td>3.3M</td>
<td>8.3</td>
<td>82M</td>
<td>74M-78M</td>
</tr>
<tr>
<td>CNNDM</td>
<td>news</td>
<td>18M</td>
<td>311K</td>
<td>39.3</td>
<td>246M</td>
<td>63-315M</td>
</tr>
</tbody>
</table>

Table 3: Data statistics: domain of text, the number of sentences, the number of paragraphs, the averaged length (number of sentences) of paragraph, the number of training instances permuted from the paragraphs, and minimum to maximum number of keywords extracted.

Paragraph datasets. Table 3 shows the paragraph datasets collected for our experiment. We collect paragraphs from various domains: the two most frequent sub-genres extracted from BookCorpus (Zhu et al., 2015) dataset: Fantasy and SciFi, Wikipedia text from wikiText-103 (Merity et al., 2016), and news articles from CNN/DailyMail (CNNDM) dataset (See et al., 2017). CNNDM and WikiText contain factual knowledge about events or things, whereas Fantasy and Romance are more narrative.

For a fair comparison, we restrict the number of sentences in a paragraph from 4 to 7, the same as the setup in Kang et al. (2019). Since CNNDM has no specific line breakers in the document, each document is regarded as a single paragraph (39.3 lengths on average). Each dataset is randomly split by 0.9/0.05/0.05 for the train, valid, and test set, respectively.

Models. As baselines, we compare non-pretrained sequence-to-sequence models: BiLSTM (Hochreiter and Schmidhuber, 1997) and hierarchical seq2seq HRED (Serban et al., 2017; Sordoni et al., 2015) by encoding the concatenation of context sentences and then decoding the target sentences. We also compare two strong paragraph generation models: FlowNet$_{disc}$ using discourse relations and FlowNet$_{latent}$ using latent delta relations (Kang et al., 2019), following the same setups (e.g., discourse parser, hyper-parameters) of the original paper.

Also, we use the pre-trained language model baselines fine-tuned on our paragraph datasets: the fine-trained bert-base-uncased (BERT$_{finetune}$) and gpt2-base (GPT2$_{finetune}$) models (Wolf et al., 2019). For BERT, we use the sequential sampling method (Wang and Cho, 2019) with Nucleus sampling strategies for producing more diverse text (Holtzman et al., 2019).

Our proposed method SSPlanner is trained using either bert-base-uncased or gpt2-base. As an upper-bound of our method, we predict masked, target text using the ground-truth plan keywords $\hat{p}$.

We find the best hyper-parameters on the validation set using a grid search on the learning rate, the number of training epochs, sampling parameters, and so on. We follow the default parameters used in the HuggingFace’s transformer models (Wolf et al., 2019). For a pointer-generator, we follow the default parameters in (See et al., 2017). The maximum number of plan keywords per sentence is 3. For more details, see the Appendix.

Metrics. We evaluate our models using both automatic metrics and human evaluation: For automatic metrics, we use two hard metrics: BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), as well as an embedding similarity metric to capture the semantic similarity: Vector Extrema (VE) (Liu et al., 2016).

For human evaluation, we measure fluency, coherence with respect to context, and overall quality with 1-5 Likert scale. We randomly select 100 samples from the test set in each Romance, WikiText, and CNNDM (total 300 paragraphs). Each sample is annotated by three crowd-workers then averaged. We also measure how human performs on the task by asking workers to predict the masked text in these 300 paragraphs.

5.1 Automatic and Human Evaluation

Table 4 and 5 show automatic and human evaluation result on PARCOM task. The fine-tuned models (BERT,GPT2)$_{finetune}$ and FlowNet models show significant improvements over the seq2seq

---

In our experiment, no fine-tuned models (original pre-trained models) show very poor performance on our task.
baselines (BiLSTM and HRED) by large margins (~1.5 METEOR), showing the importance of fine-tuning on target text and modeling inter-sentential relation, respectively.

In all datasets, SSPlanner shows significant improvements in both hard and soft metrics. This indicates that explicitly predicting content words before surface realization helps generate more coherence text on target-oriented generation in PARCOM. SSPlanner with GPT2 outperforms SSPlanner with BERT, because such autoregressive models like GPT2 are more appropriate for our task, whereas BERT is not. Finally, the performance of SSPlanner with the ground-truth keywords (\(\hat{p}\)) achieves the dramatic gain, which can be seen as an upper bound of our planning framework. Among domains, Fantasy and Romance seem to be better predicted compared to WikiText and CNNDM that require additional factual knowledge as well as narrative coherence.

Using the best model; SSPlanner (GPT2), we conduct a human evaluation on various system outputs and human-generated text (Table 5). The fine-tuned GPT2 model shows high fluency as itself but very low coherence with context, because PARCOM requires not only fluent and natural text but also context-aware text. SSPlanner achieves much higher coherence and overall quality than the baselines, but still is far behind the upper-bound model (SSPlanner with \(\hat{p}\)) and human generation.

### 5.2 Performance of Self-supervision Modules

We measure performance (i.e., accuracy) of each self-supervision module in SSPlanner: next sentence prediction (NSP) and plan prediction (PP) on the test samples (Table 6). SSPlanner achieves very high accuracy in NSP. In PP, SSPlanner correctly predicts almost half of the keywords from the total vocabulary size, indicating that the plan prediction module in SSPlanner can capture a certain level of coherence between the given context and target text to predict, although it is not perfect.

### 5.3 Comparison of Self-supervision Modules and Keyword Types in Training

Table 8 shows ablation on self-supervision modules. All scores are macro-averaged on three datasets: Romance, WikiText, and CNNDM. Each module helps improve the overall performance (METEOR): plan prediction (+2.1 M), sentiment positional embedding (+1.4 M), and next sentence prediction (+0.4 M).

Among the different types of keywords used in training (Table 9), the combination of nouns and verbs and the keywords extracted from the off-the-shelf system.
Task: given context sentences ①, ②, ③, predict target sentences ④, ⑤.

① "They reached the raised sanctuary with the slab-marble altar and the tall-backed cathedra, the bishop’s seat.” ② "Vigor and his niece made the sign of the cross.” ③ "He led them through a gate in the chancel railing.” ④ "Beyond the railing, the altar was also marked in chalk, the travertine marble stained.” ⑤ "Police tape cordoned off a section to the right.”

Plan keywords extracted from target sentences using different systems:

- **Off-the-shelf**:
  - ② (vigor, dropped, one), ③ (chancel, railing, led), ④ (travertine, marble, stained)

- **Syntactic (noun)**:
  - ② (vigor, knee), ③ (gate, chancel), ④ (railing, altar, chalk)

- **Syntactic (verb)**:
  - ② (dropped, got), ③ (led, railing), ④ (marked, stained)

- **Syntactic (nounverb)**:
  - ② (vigor, dropped, knee), ③ (led, gate, chancel), ④ (railing, altar, marked)

- **Attention**:
  - ② (vigor, dropped, got), ③ (led, gate, railing), ④ (altar, chalk, travertine)

---

**Table 7:** Example paragraph with the plan keywords extracted from different algorithms and output predictions by SSPlanner and human writer. F is fluency, C is coherence with context, and Q is overall quality.

<table>
<thead>
<tr>
<th>Models</th>
<th>M</th>
<th>VE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSPlanner</td>
<td>7.9</td>
<td>66.6</td>
</tr>
<tr>
<td>- Sentence Position (SP)</td>
<td>-1.4</td>
<td>-8.1</td>
</tr>
<tr>
<td>- Plan Prediction (PP)</td>
<td>-2.1</td>
<td>-13.2</td>
</tr>
<tr>
<td>- Next Sentence Prediction (NSP)</td>
<td>-0.5</td>
<td>-3.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models</th>
<th>M</th>
<th>VE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human writer</td>
<td>F: 4.3, C: 3.9, Q: 3.8</td>
<td></td>
</tr>
<tr>
<td>Predicted plan keywords</td>
<td>F: 4.8, C: 4.9, Q: 4.8</td>
<td></td>
</tr>
<tr>
<td>Predicted target sentences</td>
<td>F: 4.8, C: 4.9, Q: 4.8</td>
<td></td>
</tr>
</tbody>
</table>

- **SSPlanner**:
  - ④ (altar, mark, bishop)

- **Human writer**:
  - ④ (vigor, show, sanctuary), ② (gate, blood, trace), ③ (kill, sacrifice, recently)

Table 8: Ablation on self-supervision modules.

<table>
<thead>
<tr>
<th>Models</th>
<th>M</th>
<th>VE</th>
</tr>
</thead>
<tbody>
<tr>
<td>w Random</td>
<td>6.1</td>
<td>54.0</td>
</tr>
<tr>
<td>w Syntac(Verb)</td>
<td>7.6</td>
<td>63.7</td>
</tr>
<tr>
<td>w Syntac(Noun)</td>
<td>7.5</td>
<td>62.2</td>
</tr>
<tr>
<td>w Syntac(N+V)</td>
<td>8.0</td>
<td>66.3</td>
</tr>
<tr>
<td>w Off-the-shelf</td>
<td>7.8</td>
<td>66.8</td>
</tr>
</tbody>
</table>

Table 9: Comparison of plan keyword types at training (right). All scores are macro-averaged on three datasets: Romance, WikiText, and CNNDM.

shelf algorithm outperform the other types. We conjecture that since a sentence consists of both entities (i.e., nouns) and events (i.e., verbs) according to the script theory (Schank and Abelson, 2013), the combination of them provides the largest amount of information to complete the sentence. Attention-based keywords are not that helpful because the averaged attention weights themselves may not be a good indicator for topical coherence.

5.4 Comparison of Keyword Types and Ratios in Testing

In Figure 3, at test time, the predicted keywords from SSPlanner (red) shows dramatic improvements in both METEOR and VE against the random keywords (blue), but far behind the ground-truth keywords (yellow). As more predicted keywords are used at testing time, the generation quality increases. We include the full table of our ablation tests over the three datasets in the Appendix.

Table 7 shows an example paragraph with
ground-truth keywords extracted from different algorithms in ParCon and predicted target sentences with plan keywords by SSPPlaner and a human writer. In the prediction by SSPPlaner, half of the predicted keywords are used in the generation, making the story more coherent to the first two sentences and the last ending sentence. In the entire test set, we observe that about 43% of predicted keywords are actually used in generation.

6 Conclusion

A written paragraph itself contains various inductive coherence signals to be learned through self-supervision. Motivated by this, we propose a paragraph completion task for measuring textual coherence from a long document using different types of self-supervision signals. To solve the task, we propose a text planner SSPPlaner that explicitly predicts topical content keywords, and then guides the surface generator using the predicted plan keywords. SSPPlaner consists of different kinds of self-supervision modules: sentence positions, a sequence of words or sentences, and the topical relationship between context and target. Our self-supervised planning, in addition to other types of planning (e.g., discourse, goals, coreference, tenses) can be an important step toward modeling a long-term coherence in text generation.

Our results suggest several promising directions: Although our ablation tests show the effect of each self-supervision module, types of plan keywords, and the amount of keywords with respect to generation quality, there are more spaces to explore in self-supervised text planning. For example, one can study the generation quality with respect to the position of the target sentences (beginning, middle, end), the comparison of plan keywords predicted by human and system, the effect of data augmentation by their positions (e.g., masking the only middle), the generation quality with respect to the ratio between masked and unmasked sentences, and more.

Second, we can extend the set of plan keywords to be more structured like a discourse tree. For instance, one can write a simple structure like “(CAUSALITY (ELABORATE (Buy, Coffee)) (Pay, Tip, 12 dollars))” then the system can generate a long, coherent text reflected by the structure. Predicting such structural plans from context and imposing them into the generator would be a potential direction for future work.

Last, text planning is a cognitive function commonly used in human language generation. To generate more human-like utterances, different planning stages should be simultaneously combined together (Kang, 2020), such as abstractive planning, strategic planning, coherence planning, and diversity planning. Combining the heterogeneous planning systems will be a crucial step towards developing a human-like language generation.

Acknowledgements

We thank Dery Lucio, Hiroaki Hayashi, Taehyung Edvissys members at CMU, Hearst lab members at UC Berkeley, and anonymous reviewers at EMNLP 2020 for their helpful comments.

References


