DYPLOC: Dynamic Planning of Content Using Mixed Language Models for Text Generation

Xinyu Hua\textsuperscript{1}  Ashwin Sreevatsa\textsuperscript{2}  Lu Wang\textsuperscript{2}
\textsuperscript{1}Khoury College of Computer Sciences, Northeastern University, Boston, MA
\textsuperscript{2}Computer Science and Engineering, University of Michigan, Ann Arbor, MI
\textsuperscript{1}hua.x@northeastern.edu
\textsuperscript{2}\{asreeva, wangluxy\}@umich.edu

Abstract

We study the task of long-form opinion text generation, which faces at least two distinct challenges. First, existing neural generation models fall short of coherence, thus requiring efficient content planning. Second, diverse types of information are needed to guide the generator to cover both subjective and objective content. To this end, we propose DYPLOC, a generation framework that conducts dynamic planning of content while generating the output based on a novel design of mixed language models. To enrich the generation with diverse content, we further propose to use large pre-trained models to predict relevant concepts and to generate claims. We experiment with two challenging tasks on newly collected datasets: (1) argument generation with Reddit ChangeMyView, and (2) writing articles using New York Times’ Opinion section. Automatic evaluation shows that our model significantly outperforms competitive comparisons. Human judges further confirm that our generations are more coherent with richer content.

1 Introduction

Opinion articles serve as an important media to convey the authors’ values, beliefs, and stances on important societal issues. Automatically generating long-form opinion articles has the potential of facilitating various tasks, such as essay writing and speech drafting, and it is the focus of this work. Though opinion generation has been investigated for constructing arguments (Hua and Wang, 2018), writing reviews (Ni and McAuley, 2018), and producing emotional dialogue responses (Song et al., 2019), those outputs are relatively short. While impressive progress in generation has been achieved by using large pre-trained Transformers (Radford et al., 2019; Lewis et al., 2020a), directly adopting them for long-form opinion text generation poses distinct challenges.

First, large models still fall short of producing coherent text due to the lack of efficient content control and planning (Ko and Li, 2020; Wu et al., 2020; Tan et al., 2021). A common solution is to use concatenated phrases or semantic representations to guide the generation process (Yao et al., 2019; Harkous et al., 2020; Ribeiro et al., 2020; Goldfarb-Tarrant et al., 2020), where content planning, including both content selection and ordering, is expected to be learned by attention mechanisms. However, attentions have only achieved limited improvements. Recent work also explores training a
separate planning module to produce sorted content, which is then fed into a generator (Fan et al., 2019; Hua and Wang, 2020; Goldfarb-Tarrant et al., 2020). Nonetheless, this strategy results in a disconnection between planning and realization, and the output is not guaranteed to respect the planning results (Castro Ferreira et al., 2019; Prabhumoye et al., 2020).

The second challenge for opinion generation resides in the diversity of information that is needed to produce an output with consistent stances and supported by pertinent facts. Though large models memorize significant amounts of knowledge, they cannot retrieve and operate with them precisely (Lewis et al., 2020b). Due to the argumentative nature of opinion text, simply including knowledge bases (Guan et al., 2020; Zhou et al., 2020) is insufficient to uphold the desired quality, as it requires the combination of subjective claims and objective evidence as supports.

To this end, we propose a novel generation framework, DYPLOC (dynamic planning of content), to conduct content selection and ordering as text is produced. Concretely, given a set of unordered content items, as displayed in Figure 1, we design mixed language models, with each implemented as a sequence-to-sequence model to encode one item and the input statement. At each decoding step, our system selects which items to reflect, and predicts a word based on probabilities marginalized over all language models. Crucially, our end-to-end trained framework (1) enables the generator to access multiple content items at all times and select content based on what has been generated so far, (2) can be directly built on large pre-trained Transformers, e.g., BART (Lewis et al., 2020a), with planning and generation modules jointly trained, and (3) outputs learned content selection scores to provide an interface for system decision interpretation.

Furthermore, to ensure that our framework can be applied to a broad range of generation tasks, we design content items to cover three critical elements: entities and concepts that are central to many generation applications, and claims that are building blocks for opinion text. We show an example for counter-argument generation in Figure 1. Importantly, we employ BART to predict additional relevant concepts, derived from ConceptNet (Speer et al., 2017), and generate claims, as central propositions, to enrich the generated text with both objective and subjective content.

For experiments, we collect two datasets: (1) posts from Reddit ChangeMyView for argument generation, and (2) articles from the New York Times Opinion section (Sandhaus, 2008) for opinion article writing. Our proposed framework outperforms competitive comparisons, such as fine-tuning BART with the same content items, based on automatic metrics of BLEU, ROUGE, and METEOR. Human assessment further confirms that our system outputs have richer content and are more coherent in both tasks.

Our main contributions are summarized as below:

- We present a dynamic content planning generation framework, which is directly built on top of BART. Our design of mixed language models overcomes the lack of control by existing models that use implicit planning with attentions or hard copying.
- We propose content plan augmentation by automatically generating relevant concepts and claims.
- We construct two opinion text generation datasets with content plans that capture prominent entities and concepts.

2 Related Work

Neural Generation with Planning. Text planning is seen as a crucial step to guide the generation of high-quality, well-organized natural language text (McKeown, 1992; Reiter and Dale, 2000). Incorporating planning modules to neural text generator has attracted significant research interests (Shen et al., 2019; Moryossef et al., 2019; Puduppully et al., 2019), which proves to be especially beneficial for long-form output (Fan et al., 2019; Hua and Wang, 2019). More recently, large pre-trained Transformers have established new state-of-the-arts for a wide range of text generation tasks (Lewis et al., 2020a; Roller et al., 2020; Kale and Rastogi, 2020). But it is non-trivial to integrate planning modules into them. Existing approaches resort to decoupling planning and decoding stages (Hua and Wang, 2020; Kedzie and McKeown, 2020), which inevitably increases system complexities and potentially introduces cascading errors.

We take inspiration from the retrieval-augmented generation framework (Lewis et al., 2020b), which is designed to incorporate relevant documents for
question answering. Our adaptation uses a trainable plan scoring module to reflect content selection and ordering, which is more suitable for long text generation and offers better interpretability. Concurrent work by Zhang et al. (2021) presents a mixture-of-expert decoder to tackle knowledge-grounded generation. However, their score distribution for language models is fixed across all decoding steps, whereas ours is updated as generation progresses and can better reflect the dynamic nature of content planning.

Controllable Text Generation. Another related line of research investigates the controllability of generation models (Wiseman et al., 2017), including conditioning over keywords (Keskar et al., 2019; Hua and Wang, 2020; Xu et al., 2020), syntactic structures (Casas et al., 2020; Goyal and Durrett, 2020), or semantic representations (Wen et al., 2015; Elder et al., 2018). Our work differs from all previous methods as we combine different types of content, covering both objective and subjective information, and attain fine-grained sentence-level control using a novel design of mixed conditional language models.

Opinion Text Generation. Our model tackles opinion articles, which differs from traditional text generation systems that mostly concern fact-based generations (Gardent et al., 2017; Novikova et al., 2017; Puduppully et al., 2019). An extensive body of work has studied summarizing (Wang and Ling, 2016; Suhara et al., 2020; Bražinskas et al., 2020) or generating (Ni and McAuley, 2018; Li et al., 2019) reviews and building dialogue systems enhanced with emotions (Li et al., 2016; Song et al., 2019). More recently, developments are made in generating argumentative text (El Baff et al., 2019; Hidey and McKeown, 2019), which primarily focus on constructing single sentence claims on a limited number of topics. In comparison, our model can handle substantially longer output with improved quality.

3 Model

Task Formulation. Our opinion text generation framework takes as input a set of content items. Each content item consists of a title $t$, a set of entities $E_i$ 2, such as \{United_States, 9/11_attacks\}, and a set of core concepts $C_i$, such as \{attack, knowledge\}, that are often abstract notions. Our model first expands $C_i$ by predicting additional relevant concepts $C_i^+$ and optionally generates a pertinent claim $m_i$, and then outputs the final text with multiple sentences as $y = \{y_i\}$, to faithfully reflect the content items with a coherent structure. An overview of our system is illustrated in Figure 2.

Below we first describe the content item augmentation methods (§ 3.1), followed by our generator with mixed language models that condition on expanded content items (§ 3.2).

3.1 Content Item Augmentation

Concept Expansion. With limited number of entities and concepts as input, generation systems are often incapable of producing long text with rich content, resulting in hallucination (Wiseman et al., 2017; Tian et al., 2019). Therefore, from the often-abstract core concepts, we aim to predict more specific concepts that are also relevant to the given title. For instance, as displayed in Figure 1, for core concepts \{make, happen\} and

---

2Note that $i$ distinguishes the items. Their order is random.
We thus consider a concept expansion module \( g(\cdot) \), which predicts additional relevant concepts, denoted as \( C_i^+ \), by conditioning on the original content item:

\[
C_i^+ = g(t, E_i, C_i) 
\] 

While \( g(\cdot) \) can be any conditional predictor, our experiment shows that fine-tuned BART model performs best on our tasks, where it generates \( C_i^+ \) word-by-word by consuming the content item.\(^3\)

Training data construction is described in § 4.2.

**Claim Generation.** As discussed in § 1, opinion text generation should be controlled with consistent propositions, which cannot be effectively expressed by disconnected concepts. Therefore, we argue that natural languages are more suitable for delivering central claims, since they better encode stylistic languages, e.g., persuasion strategies.

Concretely, we fine-tune another BART model by taking in the title \( t \) and the entities \( E_i \), which then produces a claim with nucleus sampling for decoding (Holtzman et al., 2020). In this work, we assume the subset of content items that can be used to generate claims is known. Possible future work includes predicting such subsets and filtering claims with quality measurement.

### 3.2 Content Realization via Mixed Conditioning

After obtaining the augmented content items, we leverage the BART model to encode each of them as a sequence, as illustrated in Figure 2. Segmenter <s> is added to indicate the change of elements in a content item. Our encoders run over all items \( \{x_i\} \) in parallel, from which we extract content item representations \( \{h_i\} \), based on the last layer’s hidden states of the first token.

The standard sequence-to-sequence (seq2seq) framework models output probabilities by taking a single sequence as input. It is challenging to extend seq2seq to consider multiple sequences simultaneously, and conduct content planning concurrently. Therefore, we introduce a plan scoring network, \( d(x_i|y_{<t}) \), which learns to dynamically select and order content based on what has been produced previously while generating the outputs. As outlined in Figure 2, our generator is informed of all content items during generation. At each decoding step \( t \), the probabilities of output words are estimated as a weighted sum of all content item-conditioned language models as follows:

\[
p(y_{t}|y_{<t}) = \sum_i d(x_i|y_{<t})p(y_t|y_{<t}, x_i) 
\]

\[
d(x_i|y_{<t}) = \text{softmax}(e_{it}) 
\]

where \( p(y_{t}|y_{<t}, x_i) \) corresponds to the \( i \)-th language model with \( x_i \) as the input. Crucially, \( d(x_i|y_{<t}) \) determines the importance of \( x_i \) when generating token \( y_t \) and thus achieves the effect of content planning. We design a two-layer feed-forward network to estimate \( e_{it} \):

\[
e_{it} = W_o \tanh(W_d[h_i; s_t]) 
\]

where \( h_i \) denotes the representation of content item \( x_i, s_t \) is the decoder state, and \( W_o \) and \( W_d \) are learnable parameters. Although mixed language models have been used by Lewis et al. (2020b) to include retrieved documents for question answering, their relevance scores are given by external retrieval models, whereas our plan scorer \( d(x_i|y_{<t}) \) is learned together with the generator.

**Training and Decoding.** Our model is end-to-end trained with both the standard cross-entropy loss \( \mathcal{L}_{\text{gen}} \) over the tokens in the target generations and a separate loss \( \mathcal{L}_{\text{plan}} \) for learning \( d(x_i|y_{<t}) \):

\[
\mathcal{L}(\theta) = \mathcal{L}_{\text{gen}}(\theta) + \mathcal{L}_{\text{plan}}(\theta) 
\]

To create labels for \( \mathcal{L}_{\text{plan}} \), we leverage the correspondence between content items and target tokens, i.e., \( d(x_i|y_{<t}) \) is optimized to approach 1 if \( y_t \) is in the sentence that derives \( x_i \), otherwise 0.\(^4\) Details about training data construction is in § 4.2.

At each decoding step, the individual language models, \( p(y_{t}|y_{<t}, x_i) \), and the distribution scores, \( d(x_i|y_{<t}) \), are first calculated in parallel. We then decode each token greedily based on the mixed language models in an autoregressive way.

---

\(^3\)We also exploited a model that uses the structure of knowledge bases, e.g., ConceptNet, for learning to expand concepts, but it yields lower precision and recall than fine-tuning BART does.

\(^4\)We also experimented with a training objective consisting of the generation loss only, but the performance degraded significantly. Future directions include removing the training signals for planning.
4 Experiment Setups

We experiment with the tasks of argument generation and opinion article writing (§ 4.1). Both tasks require generating multi-sentence output, and contain a substantial amount of opinions and factual content. We describe the construction of initial content items and the training data for generating expanded concepts and claims in § 4.2. We present models for comparison in § 4.3. Finally, we provide implementation details in § 4.4.

4.1 Tasks and Datasets

Argument Generation. We collect arguments from Reddit ChangeMyView\(^5\) (CMV) community, an online forum that features argumentative discussions. Each thread begins with an original post (OP) stating an opinion towards a controversial topic, e.g., “The U.S. is too big for one government”. High-quality replies that counter-argue with the OP and are labeled with community endorsement are collected in our prior work (Hua and Wang, 2020), covering content posted from 2013 to 2018. In this work, we extend the data collection to 2019. Our goal is to generate the entire reply (i.e., the target) given the OP title. Statistics about the CMV dataset are listed in Table 1. We reserve the most recent 1,000 samples for test and another 1,000 for validation.

Opinion Article Writing. Our second task is to generate opinion articles, as collected from the New York Times (NYT) corpus (Sandhaus, 2008). We retain articles whose taxonomy labels include Top/Opinion. To ensure that articles can be processed by our computing resource, we only keep the ones with at most 20 sentences, representing 60% of all opinion articles. As shown in Table 1, NYT outputs tend to be significantly longer and contain less claims than CMV. Similarly, we keep 1,000 examples each for test and validation sets.

4.2 Content Item Construction

From target references, we describe how to automatically construct the input content items consisting of entities and core concepts, and how to collect training data to fine-tune BART to predict more specific concepts and additional claims. Prior work has demonstrated the benefits of incorporating knowledge bases for text generation (Clark et al., 2018; Puduppully et al., 2019; Guan et al., 2020). We thus consider two sources of knowledge: (1) entities from Wikipedia, which are useful for modeling events and opinion targets, and (2) concept words from ConceptNet (Speer et al., 2017), that cover more related details. Note that our setup is generally applicable to other text generation tasks, as these input items can be obtained through standard NLP pipelines, as described below.

Entity Linking. We first segment a reference into sentences. The ones with fewer than 5 tokens are discarded for content item construction. For the rest, we extract entity mentions using Stanford CoreNLP (Manning et al., 2014), and further include nominal noun phrases. For entity linking, we adopt CrossWiki (Spitkovsky and Chang, 2012), which can process our large-scale data within a reasonable amount of time. CrossWiki maps a mention to a list of frequently linked Wikipedia entries. We further manually verify and correct the linking results for the top 500 most frequent mentions.

Concept Extraction. To identify concepts in a reference, we match the lemmatized unigrams and their part-of-speech (POS) tags against all ConceptNet entries. To create a reasonably challenging task, we only keep a subset of the matches for inclusion in the core concept set (i.e., \(C_i\)), with the rest used as \(C_i^+\), to be generated by our concept expansion model. Furthermore, we conjecture that an opinion article author tends to start with high-level topics that cover more abstract topical words. We thus leverage a lexicon (Brysbaert et al., 2014) with concreteness scores, ranging from 0 (abstract) to 5 (concrete), for over 40k English words. We keep concepts that are verbs or have a concreteness score lower than 3.0. Word coverage of references by using core concepts and additionally with aug-

<table>
<thead>
<tr>
<th></th>
<th>CMV</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td># Samples</td>
<td>77,245</td>
<td>113,616</td>
</tr>
<tr>
<td>Avg. Title Len.</td>
<td>19.2</td>
<td>5.9</td>
</tr>
<tr>
<td>Avg. # Cont. Items (% w/ Claims)</td>
<td>6.8 (76.5%)</td>
<td>9.3 (38.9%)</td>
</tr>
<tr>
<td>Avg. # Core Concepts</td>
<td>3.6</td>
<td>4.8</td>
</tr>
<tr>
<td>Avg. # Predicted Concepts</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td>Avg. # Entities</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>Avg. Target Generation Len.</td>
<td>142.0</td>
<td>218.9</td>
</tr>
<tr>
<td>Cov. by Core Concepts</td>
<td>13.2%</td>
<td>14.9%</td>
</tr>
<tr>
<td>Cov. by Augmented Concepts</td>
<td>16.0%</td>
<td>18.7%</td>
</tr>
<tr>
<td>Cov. by Augmented Cont. Items</td>
<td>52.4%</td>
<td>39.1%</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the two datasets. We report average numbers of concepts and entities per content item, and the coverage of words in target generations by different input options.

\(^5\)https://www.reddit.com/r/changemyview/
mented concepts are 13.2% and 16.9% on CMV respectively, and similarly on NYT (Table 1). Finally, we train a concept generator with BART to produce $C_1^+$, conditional on $C_1$, the title, and the entities.

Claim Detection and Generation. Claims are indispensable for opinion articles. As described in §3.1, we aim to enrich content items with claims targeting the given entities within the title’s context. To this end, we first train a claim detector by fine-tuning a BERTbase (Devlin et al., 2019) sequence classifier with a dataset consisting of sentences of claims and facts. Concretely, we collect 54,802 claim sentences from Kialo⁶, a repository for debate arguments. We then sample 50,000 sentences from Wikipedia, which are treated as facts. This classifier is applied on a reference, and sentences that are labeled as claims become the target for our claim generator.

We then learn a claim generator using BART, which takes in the title and the entities, and outputs the claim. We augment our training data with replies collected from 30 active subreddits related to political discussions, with details in Appendix A. In total, 80,566 sentences, which contain at least one entity and are labeled by our classifier as claims, are kept to train the generator.

4.3 Baselines and Comparisons

We compare with three baselines: (1) RETRIEVAL first calculates the TF-IDF weighted bag-of-words vectors for each content item, which is then used to query the training set sentences. The one with the highest cosine similarity is picked for each query, which are then ordered by a trained Pointer- Network (Vinyals et al., 2015) as described in Gong et al. (2016). (2) SENTPLANNER (Hua and Wang, 2019) is an LSTM-based seq2seq model with a separate sentence planning decoder, where the planner selects keyphrases by using attentions and the generator reflects the selections. We treat our entities and concepts as keyphrases to feed to this model. (3) SEQ2SEQ is a fine-tuned BART model, whose input is the original content items without augmentation, thus does not have access to the predicted concepts and claims.

Additionally, we consider a strong comparison SEQ2SEQFULL, by fine-tuning BART with the same augmented content items as inputs as in our model. The difference is that the content items are concatenated before being used as input.

4.4 Reproducibility

We implement all models using the Huggingface Transformers library (Wolf et al., 2020) with PyTorch (Paszke et al., 2019). We use the base model for BART, which has 768 dimensional states and 6 layers for both encoder and decoder (140M parameters in total). Our newly added plan scoring network only contains 1.2M parameters, less than 1% of the pre-trained model. Our generation model is optimized using Adam (Kingma and Ba, 2014), with a batch size of 3. To improve efficiency, we adopt the mixed-precision (FP16) to train each model, using one NVIDIA Titan RTX GPU card with 24GB memory. The number of content items is limited to 10 per sample, and the numbers of entities and concepts per content item are capped at 20, respectively. We also truncate the target output to at most 200 tokens during training. Early stopping is applied over validation loss. Our model converges after being trained for 38 hours (19 epochs) on CMV, and 45 hours (15 epochs) on NYT. The best validation perplexity reaches about 6.1 after model convergence on both datasets.

5 Results

5.1 Automatic Evaluation

Here we report results on test sets with standard automatic metrics: BLEU (Papineni et al., 2002) measures the n-gram precision (here we consider up to bigrams); ROUGE (Lin, 2004), calculated based on n-gram recall; and METEOR (Denkowski and Lavie, 2014), which also accounts for synonyms. In Table 2, we first present the results when gold-standard concept expansion is used.

Our proposed DYPLOC model achieves significantly higher performance across all metrics on both datasets. In particular, the substantial lead over SEQ2SEQFULL, which has access to the same content items as ours, indicates that dynamic content planning with mixed language models produces superior generations. Among comparison models, the gap between SEQ2SEQFULL and SEQ2SEQ shows the effectiveness of content item augmentation. We also observe a significant drop for baselines without using large models, highlighting the importance of pre-training.

Ablation Study. To verify the effect of each element in content items, we further train ablated models by removing concepts, claims, or entities. The

⁶https://www.kialo.com/
where a weighted sum of content items is considered. Indeed, we find that DYPLOC's plan scores are often sharp where one content item has much higher weight than others, and in these scenarios, it is almost equivalent to the greedy selection setup.

**Results with Generated Concepts.** Table 3 lists generation results with our system generated concepts as expansion. While all systems yield worse results compared to using gold-standard concepts, our DYPLOC still outperforms other models by substantial margins, showing its robustness when input concepts are noisy. Yet it also suggests the importance of having more accurate and comprehensive concept expansion, which should be explored in the future work.

**5.2 Human Evaluation**

We hire three proficient English speakers to evaluate four key aspects of the generated outputs: (1) **grammaticality**; (2) **coherence**, measuring if the text is logical and cohesive; (3) **relevance**, gauging topic relatedness to the input title; and (4) **content richness**, assessing the specificity and whether there is enough details in the outputs. Each aspect is rated on a scale of 1 (worst) to 5 (best). In addi-

---

**Table 2: Automatic evaluation results on both tasks. We report BLEU-2, ROUGE-2, METEOR, and output length. Best scores are in bold. Our DYPLOC model statistically significantly outperforms all baselines and comparisons (randomization approximation test (Noreen, 1989), p < 0.0005).**

<table>
<thead>
<tr>
<th></th>
<th>Argument Generation (CMV)</th>
<th>Opinion Article Generation (NYT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RETRIEVAL</td>
<td>6.29</td>
<td>3.68</td>
</tr>
<tr>
<td>SENTPLANNER</td>
<td>7.78</td>
<td>3.23</td>
</tr>
<tr>
<td>Seq2SeqFull</td>
<td>29.11</td>
<td>17.71</td>
</tr>
<tr>
<td>DYPLOC (ours)</td>
<td>32.60</td>
<td>25.69</td>
</tr>
<tr>
<td>w/o All Concepts</td>
<td>7.80</td>
<td>3.68</td>
</tr>
<tr>
<td>w/o Claims</td>
<td>31.02</td>
<td>25.03</td>
</tr>
<tr>
<td>w/o Entities</td>
<td>32.11</td>
<td>25.36</td>
</tr>
<tr>
<td>Random Selection</td>
<td>12.06</td>
<td>8.25</td>
</tr>
<tr>
<td>Greedy Selection</td>
<td>32.33</td>
<td>25.60</td>
</tr>
</tbody>
</table>

**Table 3: BLEU-2 and METEOR (MTR) results on systems with predicted concepts as input. Same trends are observed on ROUGE, which are in Appendix B.**

<table>
<thead>
<tr>
<th></th>
<th>CMV</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-2</td>
<td>METEOR</td>
</tr>
<tr>
<td>RETRIEVAL</td>
<td>8.30</td>
<td>9.64</td>
</tr>
<tr>
<td>SENTPLANNER</td>
<td>7.84</td>
<td>7.76</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>18.06</td>
<td>15.96</td>
</tr>
<tr>
<td>Seq2SeqFull</td>
<td>22.84</td>
<td>17.13</td>
</tr>
<tr>
<td>DYPLOC</td>
<td>25.54</td>
<td>18.64</td>
</tr>
<tr>
<td></td>
<td>25.54</td>
<td>18.64</td>
</tr>
<tr>
<td></td>
<td>25.54</td>
<td>18.64</td>
</tr>
<tr>
<td></td>
<td>25.54</td>
<td>18.64</td>
</tr>
</tbody>
</table>

**Table 4: Human evaluation results on grammaticality (Gram.), relevance (Rel.), coherence (Coh.), and content richness (Cont.). For each sample, outputs by all three systems are ranked based on the overall preference. We show the percentage each system is ranked as the best.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CMV</td>
<td>SEQ2SEQ</td>
<td>4.19</td>
<td>3.12</td>
<td>3.19</td>
<td>2.89</td>
<td>25.1%</td>
</tr>
<tr>
<td></td>
<td>SEQ2SEQFull</td>
<td>4.24</td>
<td>3.19</td>
<td>3.23</td>
<td>3.13</td>
<td>30.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DYPLOC</td>
<td>4.26</td>
<td>3.35</td>
<td>3.35</td>
<td>3.28</td>
<td>44.7%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NYT</td>
<td>SEQ2SEQ</td>
<td>4.38</td>
<td>3.82</td>
<td>4.20</td>
<td>4.01</td>
<td>25.2%</td>
</tr>
<tr>
<td></td>
<td>SEQ2SEQFull</td>
<td>4.48</td>
<td>3.99</td>
<td>4.30</td>
<td>4.14</td>
<td>28.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DYPLOC</td>
<td>4.55</td>
<td>4.14</td>
<td>4.31</td>
<td>4.28</td>
<td>45.9%</td>
<td></td>
</tr>
</tbody>
</table>
Table 5: Sample generations on CMV [Upper] and NYT [Lower]. System generated concepts and claims are in italics. For DYLPLOC, we highlight sentence to content item alignment using colors.

<table>
<thead>
<tr>
<th>Content Items</th>
<th>DYLPLOC</th>
<th>SEQ2SEQFULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) [ENT] CO₂ [CON] death, ensuing, toll, staggering</td>
<td>Climate change denial is driven by fossil fuel industries that are paralyzing the humanities in attempting to deal with the problem. The death toll from ensuing CO₂ is staggering. There is no winning consequence for anything remotely dire that climate change denialism has. Leaving aside the issue of GM risk, simply comparing climate change denialism to climate change protesters is not going to stop anything.</td>
<td>World wide civilization is threatened with catastrophe by climate change denial. Climate change denial is driven by the fossil fuel industry, which is paralysed by the humanities. Any attempt to deal with the problem of climate change is going to cause death tolls that ensue, staggering. Leaving aside the issue of GM, the risk is simply too much to compare.</td>
</tr>
</tbody>
</table>

We randomly select 50 samples from the test sets for both tasks, and present outputs by SEQ2SEQ, SEQ2SEQFULL, and DYLPLOC in random orders. Table 4 shows that DYLPLOC receives higher scores across all aspects and tasks. In particular, the considerable differences in coherence and content richness indicate that our framework yields better content organization as well as retains more useful information. Overall, our system outputs are ranked best for 44.7% and 45.9% of the time in two tasks, significantly more than the comparisons.

Analysis on Argumentative Quality. In the ablation study, we find that our full model’s performance is similar to the version without having claims as input. We suspect this is because claims are often paraphrased or even not directly used when delivering an argument, which cannot be captured by the automatic metrics. To better understand how claims are used for generation, we randomly select 50 examples by DYLPLOC and its variant without claims, and ask the same human judges to decide whether there is a clear central argument conveyed by each generated argument on CMV.

We observe that 66.7% of the outputs by our full model are recognized as successfully delivering arguments with consistent stances, whereas only 61.3% are true for the model variant without claims. This gap confirms that claim drafts can indeed promote the argumentative quality as perceived by human readers.

6 Further Discussions

Evaluation results on generation quality have shown the effectiveness of our mixed language models. In this section, we aim to further understand the behavior of the plan scoring network, \( d(x|y_{<t}) \), such as how it affects the usage of content items for generation. Specifically, we adopt the following procedure to construct alignment between each sentence in the output and content items: for each token \( y_t \), we establish a mapping \( y_t \mapsto x_i \) if \( x_i \) is the most important item for producing \( y_t \), i.e., \( x_i = \arg\max_x d(x|y_{<t}) \), and \( d(x_i|y_{<t}) > 0.5 \). If all tokens in an entire sentence are mapped to the same \( x_i \), we consider this sentence is aligned to that content item. Based on this rule, we show sample output and corresponding alignments in Table 5.
For the rest of this section, we conduct analyses based on this alignment result. We first examine whether the model learns to utilize enough content items, i.e., high coverage. Then we provide insights on whether the generation faithfully reflects the argumentative claims using entailment relation labeling by human inspection.

**How many content items are used by the output?** Human judges have rated our model output to contain more relevant information (Table 4). We believe this can be attributed to the enhanced capacity to access and reflect the input data with dynamic content planning, as a result of mixed language models. To verify this hypothesis, we calculate the percentage of content items that are aligned to at least one output sentence. Figure 3 shows that, using our system, the coverage reaches 87.25% on CMV and 83.89% for NYT. If we replace the generated concepts with gold-standard concepts (as extracted from references) instead, the coverage exceeds 90% on both tasks. These observations indicate that our model can indeed adequately utilize the input data, with more accurate concepts further encouraging higher coverage.

**How are claim content items realized?** Claims are the central elements for opinion text construction. As mentioned in § 4.2, a subset of the content items are supplied with claim sentences. In order to examine whether they are realized as claim sentences in the outputs, we leverage the fine-tuned BERT classifier (§ 4.2) to label all output sentences. 90.96% of the sentences that are aligned to a claim element in the input are also labeled as claim on CMV. The percentage is only 69.41% for NYT, though, likely because the NYT opinion articles still contain more objective information.

Furthermore, we conduct a human evaluation study to assess the semantic relations between claim input and its aligned generated sentence. We randomly sample 50 outputs from test sets, and ask four human judges to read each. For each sample, we highlight one output sentence that is aligned to a content item with claim element. The judges determine a three-way (ENTAIL, NEUTRAL, CONTRADICTORY) entailment relation between the input claim (premise) and the output (hypothesis). Results show that ENTAIL accounts for 49.06% of all instances, while only 3.77% are deemed CONTRADICTORY. Upon inspection, the contradictory pairs are usually disagreements with regard to implicit sentiments, e.g., “Journalist is the most responsible for the problem” vs. “Media coverage is a good thing.”. This suggests that while our conditional language model achieves reasonable semantic control in most cases, it is still not guaranteed to capture more nuanced semantics encoded in opinions and arguments. Future work includes designing representations that can better model stances in opinions as well as argumentative structures.

7 Conclusion

We present a novel text generation framework that enables dynamic content planning based on mixed conditional language models. We further employ large models to augment system inputs with diverse content that covers both objective and subjective information. The experiments on two distinct opinion text generation tasks show that our proposed model compares favorably against strong comparisons based on fine-tuned BART models with the same input. Human evaluation further confirms that our model generations have richer information and better content organization.

Acknowledgements

This research is supported in part by National Science Foundation through Grant IIS-1813341. We thank three anonymous reviewers for their valuable suggestions on various aspects of this work.

Ethics Statement

Large models that are pre-trained on heterogeneous web data are shown to encode biases and can be potentially harmful for marginalized populations. Along with the improved controllability, we also recognize that our system might be misused to create fabricated or offensive content. We therefore advocate cautious and responsible practices in real-world deployment.
References


Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. Behavior research methods, 46(3):904–911.


A Training Data Construction for Claim Generator

We describe the claim generation model in § 4.2 for content item enrichment. Since both our CMV and NYT data focus on the politics domain, we leverage a collection of Reddit posts from politics related subreddits. The full list of subreddits are shown in Table 6. In total, we collect 1.6 million posts, which are split into sentences, among which we only keep the ones classified as claim by the BERTbase classifier and have at least one named entity.

<table>
<thead>
<tr>
<th>Anarchism</th>
<th>AmericanPolitics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capitalism</td>
<td>Anarcho_Capitalism</td>
</tr>
<tr>
<td>Conservative</td>
<td>democracy</td>
</tr>
<tr>
<td>democrats</td>
<td>feminisms</td>
</tr>
<tr>
<td>government</td>
<td>GreenParty</td>
</tr>
<tr>
<td>IWW</td>
<td>labor</td>
</tr>
<tr>
<td>Liberal</td>
<td>Libertarian</td>
</tr>
<tr>
<td>LibertarianLeft</td>
<td>LibertarianSocialism</td>
</tr>
<tr>
<td>Marxism</td>
<td>moderatepolitics</td>
</tr>
<tr>
<td>Objectivism</td>
<td>PoliticalDiscussion</td>
</tr>
<tr>
<td>politics</td>
<td>progressive</td>
</tr>
<tr>
<td>Republican</td>
<td>republicans</td>
</tr>
<tr>
<td>socialdemocracy</td>
<td>socialism</td>
</tr>
<tr>
<td>ukpolitics</td>
<td>uspolitics</td>
</tr>
<tr>
<td>worldpolitics</td>
<td>PoliticalPhilosophy</td>
</tr>
</tbody>
</table>

Table 6: List of subreddits used to construct training data for learning the claim generator.

<table>
<thead>
<tr>
<th>CMV</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td>RETRIEVAL</td>
<td>ROUGE-2</td>
</tr>
<tr>
<td>4.39</td>
<td>82</td>
</tr>
<tr>
<td>SENTPLANNER</td>
<td>3.24</td>
</tr>
<tr>
<td>SEQ2SEQFULL</td>
<td>8.83</td>
</tr>
<tr>
<td>DYPLOC</td>
<td>11.83</td>
</tr>
</tbody>
</table>

Table 7: ROUGE-2 and average length (Len.) on systems with predicted concepts as input.

B Additional Automatic Evaluation Results

In § 5.1, we report results by automatic metrics using system predicted concepts in Table 3. Here we additionally show the results evaluated by ROUGE-2 and average output lengths in Table 7.

C Human Evaluation Guideline

We include the detailed human evaluation guidelines in Figure 4. Note that we collect 53 samples for annotation for each domain. The first three are for calibration only and not be included in the final results.

D Additional Sample Outputs

Additional example content items and generations are demonstrated in Table 8 and Table 9.
In the following studies, you will evaluate the system outputs of three text generation models on two different domains. For each domain, there will be 53 examples presented, each starting with a statement, followed by three system generations. Please first read the statement and then the system outputs. At the end of each output, please provide your judgment on the quality of the following aspects, based on a scale of 1 (worst) to 5 (best):

- **Grammaticality**: whether the text reads fluently and has no grammar error
  - 1. Major grammatical errors that significantly impact comprehension of text. E.g., “I’m not a quick skimming, but I’m quickly making a comment.”
  - 3. Minor grammatical errors that do not significantly impact comprehension of text. E.g., “I have car that works, and I make it to work by commute 45 minutes to an hour on my bike.”
  - 5. No grammatical issues. E.g., “There are swathes of people whose function is determined by technology, and they use technology as a crutch.”

- **Coherence**: whether the information transition is natural and well-structured
  - 1. Sentences are completely unrelated. E.g., “The Supreme Court created a mechanism for interpreting the Constitution through a modern lens. The question is, do you create jobs? Ukraine is a direct ally of the US.”
  - 3. Sentences are connected to one another but transitions seem disjointed; there doesn’t appear to be a strong progression of ideas. E.g., “Muslims worship the figure of Allah. Christians worship the figures of God. Muslims do not worship the Jews. Muslims don’t worship the Christian figure of God, Muslims worship God. They worship the Jewish figure of the figure.”
  - 5. Sentences transition smoothly and connect to form a logical progression. E.g., “Every country has to deal with their own geography. USA benefits from decent climate country wide, plentiful natural resources and distance from areas of war. The downside is that they are close to Mexico and Mexico pretty much sucks, so it’s inhabitants want to get into the USA. Unless you believe that all resources and other benefits should be shared then why should the world take on the USA downsfalls while not getting any of the plusses?”

- **Relevance**: whether the content of the text is relevant to the title
  - 1. The output is generic or completely irrelevant. E.g., “Supply and demand. The US thinks those drugs are worth price X. Other countries are only willing to pay price Y. The US develops more IP related content than other countries because it has a huge military and is able to enforce IP laws.”
  - 3. The text is tangential to the title and the input (it may share an entity or key concept in common), though it might not be precisely on topic. E.g., “When you enter a college career, you decide to take literature studies. You can become an engineer, history, linguistics, etc.”
  - 5. The text is highly relevant with the title and the input. E.g., “The problem with protesting minority opinions is that you force the majority opinion to come out against them, and as a result you find controversial speakers turning their commencement speeches into bland speeches. Commencement speeches are a recognition of a person, and offer an affirmation of their worldview.”

- **Content Richness**: whether the output conveys substantial content
  - 1. Generic response with no useful information about the topic. E.g., “I don’t have time to address the point you’re making.”
  - 3. With one or two key points that are useful as counter-argument. E.g., “Reducing costs is not the goal of the free market.”
  - 5. With sufficient key information that is useful as counter-argument. E.g., “Reducing costs is not the goal of the free market. Simply setting prices for medical procedures has been shown to be extremely effective. I will tacitly admit that your post is true for many countries, but the US health share is less than 1% of GDP.”

Figure 4: Human evaluation guidelines and representative examples on rating scales.
<table>
<thead>
<tr>
<th>Title</th>
<th>I feel that humanity is going backwards due to a rise in internet use and lack of face to face communication.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td><a href="https://www.reddit.com/r/changemyview/comments/1xklhm">https://www.reddit.com/r/changemyview/comments/1xklhm</a></td>
</tr>
</tbody>
</table>

**Content Items**

(a) [CON] give, coherent, impossible, argue, argument, view (b) [CON] intent, normal, face, conversation, interpret, body, language, facial, gesture, pitch, tone, content (c) [ENT] Internet [CON] true, relate, assertion, change, advent, stop, move, forward [CLAIM] I do believe that the internet has been beneficial in the past for communication, but I believe in a society where we are constantly moving forward and we can always adapt (d) [CON] modern, affect, attitude, communication, generation, current, future, society, skill, relationship

**DYLOP**

I don’t think modern society is affecting attitudes, communication skills, or relationships in current and future generations. I think the intent of normal face to face conversations is to interpret body language, facial gestures, pitch tone, content, etc. This is true for all communication related to the internet. Your assertion that the changes caused by the advent of the internet stop us from moving forward. You haven’t given a coherent argument for your view, and it is impossible to argue against it.

**SEQ2SEQFULL**

Modern society is affected by the attitudes of communication skills and relationships of current and future generations. The intent of normal face to face conversation is to interpret body language, facial gestures, pitch tone, and content. This is true for all communication related to the internet. Your assertion that society is changing because of the advent of the internet is not true. It’s not that society has stopped moving forward, it’s that society hasn’t moved forward. I can’t give a coherent argument for your view, because it’s impossible to argue against it.

<table>
<thead>
<tr>
<th>Title</th>
<th>I believe all police in the UK should be armed with both tasers and guns. CMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td><a href="https://www.reddit.com/r/changemyview/comments/1z4s02">https://www.reddit.com/r/changemyview/comments/1z4s02</a></td>
</tr>
</tbody>
</table>

**Content Items**

(a) [ENT] Police, Alternative_rock [CON] function, prevent, idea, force, literally, crime, military, founding (b) [ENT] Gun [CON] comment, argue, agree, important, aspect, consent, principles (c) [ENT] Police, Filename_extension [CON] separate, law, completely, power, elevate, people (d) [CON] protect, group, people, pay, full, time, citizen, responsibility, basically

**DYLOP**

The founding idea of the police force was literally to function as an alternative to preventing crime, and as such, as a military force. Basically, a group of people who are paid full time as US citizens have a responsibility to protect themselves. Separating the police from the law completely elevates the power of the people. A lot of the comments argue that guns are necessary, but I agree that the most important aspect of police consent is the principles of consent.

**SEQ2SEQFULL**

I see a lot of comments arguing for guns, and I agree with you that the most important aspect of police is consent. The principles of the founding of the idea of the police force are literally the only ones that function in preventing crime. The military is basically an extension of the Police force. It’s a group of people who pay full time citizens to take responsibility for protecting themselves. Separating them from the law completely elevates their power.

<table>
<thead>
<tr>
<th>Title</th>
<th>I do not believe democracy inherently protects freedom. CMV.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td><a href="https://www.reddit.com/r/changemyview/comments/1yg8sx">https://www.reddit.com/r/changemyview/comments/1yg8sx</a></td>
</tr>
</tbody>
</table>

**Content Items**

(a) [ENT] Democracy, Democratic_Party_(United_States) [CON] respects, grow, modern, freer, stride, freedom, century, democratic (b) [CON] modern, democracy, job, protect, freedom, government, people

**DYLOP**

Modern democracies do a better job of protecting people’s freedom from government. In that respect, modern democracies have grown freer over time, and the strides in freedom that took place in the 20th century were mostly because of the democratic movement.

**SEQ2SEQFULL**

Modern democracy’s job is to protect people’s freedom. The government in many respects has grown freer over time, and has taken on a different striding place in the 20th century. The democratic movement was formed during the time of the time, during the 2020th century, and during the 21st century, the democratic movement has taken over. The movement has done a better job of protecting freedom than modern democracy.

Table 8: Sample content items and system output on CMV. System generated concepts and claims are in *italics.*
To the Editor: The irony is apparent: the health of the American economy may end up hurting the world, as you point out in your Dec. 9 editorial “Rebounding Wall Street.” Multinational agreements like the Investment and Important Trade Treaty negotiated in Paris, with the full support of the United States, will increase capital mobility, pointing investors to developed nations and smaller countries with the throat of the global economy. In addition to limiting rights imposed by environmental restrictions on foreign investors, including ownership of natural resources and media, the treaty will have a cumulative effect on places like traditional national parks and tourist attractions. Tearing down national barriers to capital flight will lead to the kind of starvation that is a key essential tenet of trade policy.

To the Editor: The irony is apparent: the health of the American economy will not end up hurting the world as you point out (editorial, Dec. 17). The rebound from Wall Street has torn down national barriers to capital flight and led to a kind of starvation. A key essential tenet of American trade policy is the Multinational Agreement on Investment. This important treaty was negotiated in Paris with the full support of the United States. It increased the mobility of investors from developed nations to smaller countries, and it has helped keep the United States out of the world’s throat. In addition, it limits rights of nations to impose environmental restrictions on foreign investment, including ownership of natural resources and media rights.

To the Editor: Re “Funds Big and Fail, Fed Assists in Bailout” (front page, Feb. 13): It is difficult to understand why the Federal Reserve Bank of New York should not organize a privately held hedge fund for long-term investment. Federal regulators arranged the bailout because the market didn’t work. This country has asserted the efficiency of free-market forces and has consistently advocated free-market principles to the global community. MICHAEL J. BERMAN New York, Feb. 13, 1999

The writer is vice chairman of fiduciary trust at Prudential International.

To the Editor: Re “Funds Big and Fail, Fed Assists in Bailout” (front page, Dec. 17): It is difficult to understand why the Federal Reserve Bank of New York should not organize a privately held hedge fund to manage long-term capital. The Federal Reserve has been the only central bank of central bank management and regulatory governance in the country to assert the efficiency of free-market forces. The Fed has consistently advocated the principle that the global community should not allow regulators to arrange the work of hedge funds. MICHAEL J. KAPLAN New York, Dec.” 17, 1998

The writer is vice chairman of the fiduciary trust at the International Monetary Fund.

Table 9: Sample content items and system output on NYT. System generated concepts and claims are in italics.