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

Xinyu Hua[1] **Ashwin Sreevatsa**[2] **Lu Wang**[2]
[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA
[2]Computer Science and Engineering, University of Michigan, Ann Arbor, MI
[1]hua.x@northeastern.edu
[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.

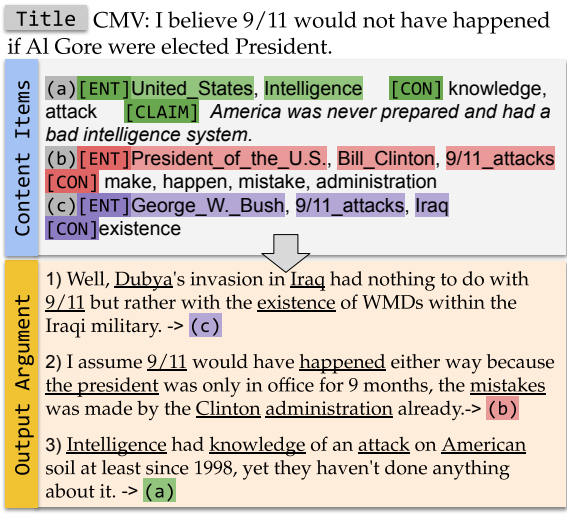

Figure 1: Sample counter-argument on Reddit ChangeMyView. Our generator considers an input containing (1) a title and (2) an unordered set of content items. Each content item consists of elements of an *entity set* [ENT], a *concept set* [CON], and an optional one-sentence *claim* [CLAIM]. Each output token is generated by conditioning on all content items, and the best aligned ones (learned by our model) are highlighted in corresponding colors. We also underline words that reflect the input concepts and entities.

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

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6408–6423
August 1–6, 2021.

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.[1] 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

[1] Data and code are available at: `xinyuhua.github.io/Resources/acl21/`.

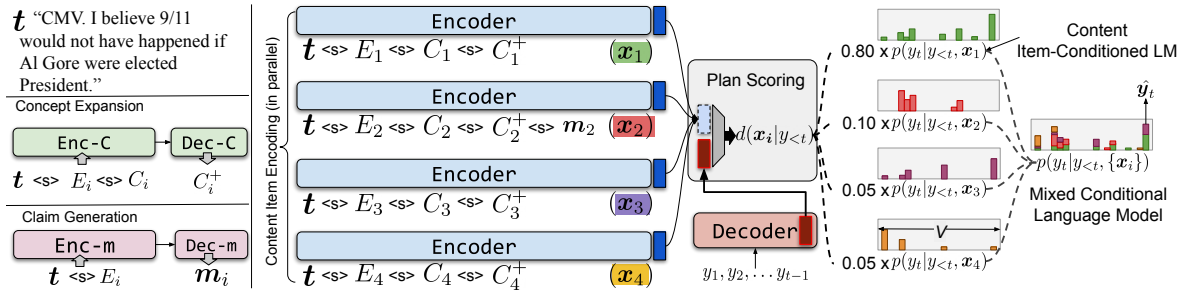

Figure 2: Our proposed text generation framework, DYPLOC. [Left] For each input content item (a title $\boldsymbol{t}$, an entity set E_i, and a core concept set C_i), we first expand it with more relevant concepts, i.e., C_i^+. For sentences to be realized as claims, we employ a separate generator to produce one draft claim, $\boldsymbol{m}_i$. [Right] The augmented content items, denoted as $\{\boldsymbol{x}_i\}$, are encoded in parallel. At each decoding step, a plan scoring network estimates a distribution $d(\boldsymbol{x}_i|y_{<t})$ for all content items and decides on relevant content. A word is predicted based on probabilities marginalized over all content item-conditioned language models, i.e., $p(y_t|y_{<t}, \boldsymbol{x}_i)$ for the i-th model.

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 $\boldsymbol{t}$, 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** $\boldsymbol{m}_i$, and then outputs the final text with multiple sentences as $\boldsymbol{y} = \{y_t\}$, 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

[2]Note that i distinguishes the items. Their order is random.

entities {Bill_Clinton, 9/11_attacks}, we grow the input with more concrete concepts of {*mistake*, *administration*}.

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:

$$C_i^+ = g(\boldsymbol{t}, E_i, C_i) \tag{1}$$

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 $\boldsymbol{t}$ 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 $\{\boldsymbol{x}_i\}$ in parallel, from which we extract content item representations $\{\boldsymbol{h}_i\}$, 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**, $d(\boldsymbol{x}_i|y_{<t})$, 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:

$$p(y_t|y_{<t}) = \sum_i d(\boldsymbol{x}_i|y_{<t}) p(y_t|y_{<t}, \boldsymbol{x}_i) \tag{2}$$

$$d(\boldsymbol{x}_i|y_{<t}) = \text{softmax}_i(e_{it}) \tag{3}$$

where $p(y_t|y_{<t}, \boldsymbol{x}_i)$ corresponds to the i-th language model with $\boldsymbol{x}_i$ as the input. Crucially, $d(\boldsymbol{x}_i|y_{<t})$ determines the importance of $\boldsymbol{x}_i$ 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}:

$$e_{it} = \boldsymbol{W}_o \tanh(\boldsymbol{W}_d[\boldsymbol{h}_i; \boldsymbol{s}_t]) \tag{4}$$

where $\boldsymbol{h}_i$ denotes the representation of content item $\boldsymbol{x}_i$, $\boldsymbol{s}_t$ is the decoder state, and $\boldsymbol{W}_o$ and $\boldsymbol{W}_d$ 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 $d(\boldsymbol{x}_i|y_{<t})$ is learned together with the generator.

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

$$\mathcal{L}(\theta) = \mathcal{L}_{gen}(\theta) + \mathcal{L}_{plan}(\theta) \tag{5}$$

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

At each decoding step, the individual language models, $p(y_t|y_{<t}, \boldsymbol{x}_i)$, and the distribution scores, $d(\boldsymbol{x}_i|y_{<t})$, 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.

	CMV	NYT
# Samples	77,245	113,616
Avg. Title Len.	19.2	5.9
Avg. # Cont. Items (% w/ Claims)	6.8 (76.5%)	9.3 (38.9%)
Avg. # Core Concepts	3.6	4.8
Avg. # Predicted Concepts	4.2	4.3
Avg. # Entities	0.8	0.7
Avg. Target Generation Len.	142.0	218.9
Cov. by Core Concepts	13.2%	14.9%
Cov. by Augmented Concepts	16.9%	18.7%
Cov. by Augmented Cont. Items	52.4%	39.1%

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.

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-

[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_i^+, conditional on C_i, 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 $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) sequence classifier with a dataset consisting of sentences of `claims` and `facts`. Concretely, we collect 54,802 claim sentences from Kialo[6], 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.

[6] https://www.kialo.com/

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

	Argument Generation (CMV)				Opinion Article Generation (NYT)			
	BLEU-2	**ROUGE-2**	**METEOR**	**Len.**	**BLEU-2**	**ROUGE-2**	**METEOR**	**Len.**
RETRIEVAL	6.29	3.68	10.00	78	9.68	7.96	9.98	99
SENTPLANNER	7.78	3.23	7.69	114	7.45	5.06	6.62	106
SEQ2SEQ	16.71	9.53	13.34	100	21.44	14.92	14.93	119
SEQ2SEQFULL	29.11	17.71	20.27	145	31.06	29.74	23.10	121
DYPLOC (ours)	**32.60**	**25.69**	**22.61**	101	**40.63**	**36.93**	25.76	122
w/o All Concepts	7.80	3.68	7.21	107	11.32	6.01	8.33	132
w/o Augmented Concepts	22.39	15.90	16.91	99	26.94	21.56	18.39	117
w/o Claims	31.62	25.03	22.09	100	39.44	35.43	25.25	122
w/o Entities	32.11	25.36	22.42	101	39.66	35.82	25.11	122
Random Selection	12.96	8.25	10.05	103	5.32	5.29	6.00	72
Greedy Selection	32.33	25.60	22.53	100	40.61	36.88	**25.77**	122

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), $p < 0.0005$).

	CMV		NYT	
	BLEU-2	**METEOR**	**BLEU-2**	**METEOR**
RETRIEVAL	8.30	9.64	8.85	9.21
SENTPLANNER	7.84	7.76	7.75	6.80
SEQ2SEQFULL	18.06	15.96	16.20	15.25
DYPLOC	**22.84**	**17.13**	**24.54**	**17.41**

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.

Data	**System**	**Gram.**	**Coh.**	**Rel.**	**Cont.**	**Top-1**
CMV	SEQ2SEQ	4.19	3.12	3.19	2.89	25.1%
	SEQ2SEQFULL	4.24	3.19	3.23	3.13	30.2%
	DYPLOC	**4.26**	**3.35**	**3.35**	**3.28**	**44.7%**
NYT	SEQ2SEQ	4.38	3.82	4.20	4.01	25.2%
	SEQ2SEQFULL	4.48	3.99	4.30	4.14	28.9%
	DYPLOC	**4.55**	**4.14**	**4.31**	**4.28**	**45.9%**

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.

results are also displayed in Table 2. In general, scores decrease when using only partial content items, among which removing all concepts lead to the biggest performance drop, suggesting that entities and claims alone are insufficient to produce informative outputs.

Effect of Hard Selection of Content Items. To test the necessity of using weighted-sum marginalization (Eq. 2), we experiment with two comparisons with hard selections, i.e., either randomly choosing a content item, or using the one with the highest predicted plan score (greedy selection). For both cases, we set the selected content item's plan score as 1.0, with the rest of the candidates having a score of 0.0, to ensure the probabilities summed up to 1.0. As can be seen from the bottom two rows of Table 2, not surprisingly, random selection performs much worse. We observe that its generations lack coherence and fluency, implying the effectiveness of our learnable content planner. On the other hand, using greedily selected content items obtains comparable results with DYPLOC, 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-

Content Items	(a) [ENT] CO2 [CON] death, ensue, *toll*, *staggering* (b) [CON] leave, stop, compare, change, denialism, issue, *simply*, *risk* (c) [ENT] Fossil_fuel [CON] drive, paralyze, deal, *humanity*, *industry* [CLAIM] *Coal is not a reliable source of energy, and it's been driven by unreliable sources of energy and unreliable sources to date.* (d) [CON] win, consequence, remotely, *dire*
DYPLOC	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.
SEQ2SEQFULL	World wide civilisation is threatened with catastrophe by climate change denial. Climate change denial is driven by the fossil fuel industry, which is paralyseed 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 is simply too much to compare.
Content Items	(a) [ENT] President_of_the_U.S., George_W._Bush, Richard_Nixon [CON] omit, *list* (b) [ENT] George_W._Bush, [CON] bring, people, world, *honor*, *dignity*, *respect* (c) [ENT] Bill_Clinton, Ronald_Reagan [CON] harm, respect, demonstrate, damage, *office*, *state* (d) [CON] owe, correction, omission, *apology*, *readers*
DYPLOC	President Clinton "has done more harm to the respect of the Presidency than Presidents Presidents Ford, Carter and Reagan have demonstrated," and that "the failure of Presidents Clinton and Reagan has caused permanent damage to the office." You omit President Bush from your list of Presidents. Mr. Bush has brought honor, dignity and respect to the office of the Presidency in the eyes of the American people and the world at large. We owe your readers an apology and a correction of your editorial's glaring omission, which created an erroneous impression.
SEQ2SEQFULL	"The Picture of Bill Clinton" (editorial, Dec. 5) states that President Clinton "has done more harm to our respect for the Presidency than Presidents Reagan, Ford, Carter and Bush." This demonstrates that President Reagan's failure to do more damage to our honor than President Bush's failure in office. You omitted from your list President Clinton's achievements that brought honor and dignity to the eyes of the American people and to the world at large. [...]

Table 5: Sample generations on CMV [Upper] and NYT [Lower]. System generated concepts and claims are in *italics*. For DYPLOC, we highlight sentence to content item alignment using colors.

tion, judges also rank the system outputs by their overall preferences. Detailed evaluation guideline is attached in Appendix C.

We randomly select 50 samples from the test sets for both tasks, and present outputs by SEQ2SEQ, SEQ2SEQFULL, and DYPLOC in random orders. Table 4 shows that DYPLOC 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 DYPLOC 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, $d(\boldsymbol{x}|y_{<t})$, 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 $y_t \mapsto \boldsymbol{x}_i$ if $\boldsymbol{x}_i$ is the most important item for producing y_t, i.e., $\boldsymbol{x}_i = \text{argmax}_{\boldsymbol{x}}\, d(\boldsymbol{x}|y_{<t})$, and $d(\boldsymbol{x}_i|y_{<t}) > 0.5$. If all tokens in an entire sentence are mapped to the same $\boldsymbol{x}_i$, we consider this sentence is aligned to that content item. Based on this rule, we show sample output and corresponding alignments in Table 5.

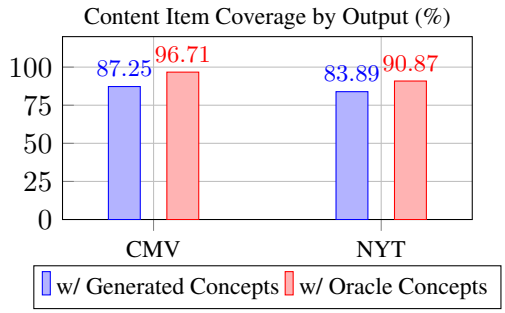

Figure 3: The percentage of content items that are aligned to at least one output sentence.

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

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

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.

Noe Casas, José A. R. Fonollosa, and Marta R. Costa-jussà. 2020. Syntax-driven iterative expansion language models for controllable text generation. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 1–10, Online. Association for Computational Linguistics.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Henry Elder, Sebastian Gehrmann, Alexander O'Connor, and Qun Liu. 2018. E2E NLG challenge submission: Towards controllable generation of diverse natural language. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 457–462, Tilburg University, The Netherlands. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252, Online. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1756–1767, Minneapolis, Minnesota. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 591–602, Hong Kong, China. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Chris Kedzie and Kathleen McKeown. 2020. Controllable meaning representation to text generation: Linearization and data augmentation strategies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5160–5185, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing discourse relations in language generation from GPT-2. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 52–59, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Junyi Li, Wayne Xin Zhao, Ji-Rong Wen, and Yang Song. 2019. Generating long and informative reviews with aspect-aware coarse-to-fine decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1969–1979, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Kathleen McKeown. 1992. *Text generation*. Cambridge University Press.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590, Hong Kong, China. Association for Computational Linguistics.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuan-Jing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2692–2700. Curran Associates, Inc.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2020. A controllable model of grounded response generation. *arXiv preprint arXiv:2005.00613*.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2021. Joint retrieval and generation training for grounded text generation. *arXiv preprint arXiv:2105.06597*.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1006–1014.

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 $\text{BERT}_{\text{base}}$ classifier and have at least one named entity.

Anarchism	AmericanPolitics
Capitalism	Anarcho_Capitalism
Conservative	democracy
democrats	feminisms
government	GreenParty
IWW	labor
Liberal	Libertarian
LibertarianLeft	LibertarianSocialism
Marxism	moderatepolitics
Objectivism	PoliticalDiscussion
politics	progressive
Republican	republicans
socialdemocracy	socialism
ukpolitics	uspolitics
worldpolitics	PoliticalPhilosophy

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

	CMV		NYT	
	ROUGE-2	**Len.**	**ROUGE-2**	**Len.**
RETRIEVAL	4.39	82	6.64	95
SENTPLANNER	3.24	115	5.12	108
SEQ2SEQFULL	8.83	120	8.83	135
DYPLOC	**11.83**	118	**15.46**	134

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 im 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 downfalls while not getting any of the plusses?"*
- **Relevance:** whether the content of the text is relevant to the title
  Title: *The recent swell in protesting Commencement speakers at colleges is a good thing.*
  - 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.

Title	I feel that humanity is going backwards due to a rise in internet use and lack of face to face communication.
Link	https://www.reddit.com/r/changemyview/comments/1xklhm
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*
DYPLOC	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, but it is not related to 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 and tone content. This is true of 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.
Title	I believe all police in the UK should be armed with both tasers and guns. CMV
Link	https://www.reddit.com/r/changemyview/comments/1z4s02
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*
DYPLOC	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
Title	I do not believe democracy inherently protects freedom. CMV.
Link	https://www.reddit.com/r/changemyview/comments/1yg8sx
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*
DYPLOC	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*.

Title	Can the Global Economy Really Stabilize Itself?
Content Items	(a) [ENT] Nation [CON] include, impose, limit, foreign, ownership, natural, cumulative, *investor*, *treaty*, *media*, *traditional* [CLAIM] *If the nation is already wealthy and can't afford to spend its money on its own people, why should it spend their money on other people?* (b) [ENT] United_States [CON] support, increase, developed, investments, multinational, agreement, capital, *treaty*, *country*, *investor*, *negotiate* [CLAIM] *America is not a very good example of that, but it does seem like we are in the process of stabilizing.* (c) [ENT] United_States [CON] hurt, rebound, irony, economy, health, *editor*, *street*, *wall* [CLAIM] *America is not a very good example of that, but it does seem like we are in the process of stabilizing.* (d) [CON] tear, national, barrier, capital, flight, lead, kind, starvation, key, essential, tenet, trade, policy
DYPLOC	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.
SEQ2SEQFULL	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.
Title	Why Save Risky Hedge Fund?
Content Items	(a) [ENT] Chairman [CON] trust, vice, *international* (b) [ENT] Box_office_bomb [CON] bailout, fail, understand, organize, hold, hedge, capital, difficult, federal, *bank*, *governance*, *regulatory*, *term* (c) [CON] advocate, force, assert, efficiency, free, consistently, principle, *global*, *market*, *community*, *country* (d) [CON] bailout, arrange, federal, *regulator*, *market*
DYPLOC	To the Editor: Re "Funds Big and Fail, Fed Assists 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.
SEQ2SEQFULL	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*.