AEG: Argumentative Essay Generation via A Dual-Decoder Model with Content Planning

Jianzhu Bao[1,4]*, **Yasheng Wang**[2], **Yitong Li**[2,3], **Fei Mi**[2], **Ruifeng Xu**[1,4,5]†
[1]Harbin Insitute of Technology, Shenzhen, China
[2]Huawei Noah's Ark Lab
[3]Huawei Technologies Co., Ltd.
[4]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[5]Peng Cheng Laboratory, Shenzhen, China
jianzhubao@gmail.com, xuruifeng@hit.edu.cn
{wangyasheng, feimi2, liyitong3}@huawei.com

Abstract

Argument generation is an important but challenging task in computational argumentation. Existing studies have mainly focused on generating individual short arguments, while research on generating long and coherent argumentative essays is still under-explored. In this paper, we propose a new task, Argumentative Essay Generation (AEG). Given a writing prompt, the goal of AEG is to automatically generate an argumentative essay with strong persuasiveness. We construct a large-scale dataset, ArgEssay, for this new task and establish a strong model based on a dual-decoder Transformer architecture. Our proposed model contains two decoders, a planning decoder (PD) and a writing decoder (WD), where PD is used to generate a sequence for essay content planning and WD incorporates the planning information to write an essay. Further, we pre-train this model on a large news dataset to enhance the plan-and-write paradigm. Automatic and human evaluation results show that our model can generate more coherent and persuasive essays with higher diversity and less repetition compared to several baselines.[1]

1 Introduction

Automatic argument generation, literately the task of generating persuasive arguments on controversial issues (Toulmin, 2003; Zukerman et al., 2000), has received many research interests in recent years (Khatib et al., 2021; Schiller et al., 2021). Many works have involved different arguments generation such as the counter-arguments generation (Hua and Wang, 2018; Hua et al., 2019; Hidey and McKeown, 2019; Alshomary et al., 2021b) and the controlled arguments generation under certain topics or aspects (Gretz et al., 2020; Schiller et al., 2021; Alshomary et al., 2021a; Khatib et al., 2021). However, real-life scenarios like news editorials, competitive debating, and even television shows, are requiring more powerful ways of systematically organizing arguments in composing long-form essays or speeches that can fully express opinions and persuade the audiences. Previous studies predominantly focused on generating individual and rela-

Writing Prompt:
Online education is becoming more and more popular. Some people claim that e-learning has so many benefits that it will replace face-to-face education soon. Others say that traditional education is irreplaceable. Discuss both views and give your opinion.

Argumentative Essay:
Acquiring knowledge virtually has become extremely popular in the present times. While many individuals believe that there are various advantages and might overtake traditional learning in the future, a sizeable group thinks that the traditional method cannot be replaced. I believe that the use of the classroom might reduce, but it cannot be replaced. This essay will discuss both views and substantiate my view in the course of the essay.

To commence with, virtual learning is widely implemented because it is convenient and cost-effective. It provides us with the opportunity to obtain an education without the hassle of travelling. Students, for instance, can attend classes at the comfort of their home, resulting in saving time that might have been spent on commuting in the past. Similarly, learning online can also be considered cost-efficient. Instead of spending an immense amount of college funds, we can attain the same level of qualifications at a cheaper price as it does not involve infrastructure.

On the contrary, traditional learning offers guided learning and hands-on experience. Classroom teaching practices assist students in obtaining better study skills, such as organising and gathering reliable information due to constant interaction with the teacher, resulting in improved academic achievement. In addition, it helps in gaining practical knowledge through sessions in laboratories that are not a part of digital practices. For example, pupils are provided with constant guidance during face to face teaching along with acquiring real-time experience.

In conclusion, although online classes might seem beneficial in terms of convenience together with being budget-friendly, classroom education provides a better learning and practical experience. Therefore, I think that face-to-face classes are not replaceable.

Table 1: An example of our proposed Argumentative Essay Generation task. Given a writing prompt about a controversial topic, the task is to generate a well-organized argumentative essay with nice coherence and strong persuasiveness. The major claims express the topic, stance, and main idea of this essay.

* This work was done during the internship at Huawei Noah's Ark Lab.
† Corresponding Author
[1]Code and data are available at https://github.com/HITSZ-HLT/AEG

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5134–5148
December 7-11, 2022

tively short arguments, which can be weak when addressing these long-form argument generation tasks.

In this paper, we aim with the question of how to generate and compose a comprehensive and coherent argumentative essay, which can contain multiple arguments with different aspects. This is a challenging but fundamental task, requiring much more capability of understanding human intelligence towards general artificial intelligence to fully address this problem (Slonim et al., 2021). However, with superior development of pre-training methods (Devlin et al., 2019; Brown et al., 2020; Bommasani et al., 2021), generating coherent long-form documents is touchable with reasonable qualities (Guan et al., 2021; Yu et al., 2021). Therefore, to facilitate this line of research, we introduce a new document-level generation task, Argumentative Essay Generation (AEG), which focuses on generating long-form argumentative essays with strong persuasiveness given the writing prompt. An example of AEG is shown in Table 1. In this example, the given writing prompt specifies a topic about "*online education*". The expected argumentative essay first introduces the topic and states the stance (paragraph 1), then justifies its point through a series of arguments (paragraphs 2-3), and finally summarizes and echos the main idea (paragraph 4). We can see that AEG requires generating relevant claims and evidences of diverse aspects relevant to a given topic, and further appropriately incorporating them in a logical manner to compose an argumentative essay.

In order to make progress towards AEG, we construct a large-scale dataset, ArgEssay, containing 11k high-quality argumentative essays along with their corresponding writing prompts on a number of common controversial topics such as technological progress, educational methodology, environmental issues, etc. Our proposed dataset is built upon the writing task of several international standardized tests of English, such as IELTS and TOEFL, which also being studied in other tasks of automated essay scoring (Blanchard et al., 2013) and argument mining (Stab and Gurevych, 2017). Compared to previous argument generation datasets collected from social media, the essays in our dataset are more formal in wording and writing and therefore of higher quality, making our dataset a better choice for studying argument generation.

To tackle the proposed AEG task, we adopt the plan-and-write paradigm for generating diverse and content-rich argumentative essays, as content planning proves to be beneficial for long-form text generation (Fan et al., 2019; Hua and Wang, 2019). We establish encoder-decoder based Transformer models with dual-decoder, which contains a planning decoder (PD) for generating keywords or relational triplets as essay content planning and a writing decoder (WD) for composing an essay guided by the planning. Adopting this dual-decoder architecture can keep planning and writing process separate to avoid mutual interference. Automatic evaluation results show that our model outperforms several strong baselines in terms of diversity and repetition. Human evaluation results further demonstrate that the essays generated by our model maintain good coherence and strong persuasiveness. We also show that our model yields better plannings compared to baselines, and the content of the generated essays can be effectively controlled by the plannings. In addition, the performance of our model can be further improved after being pre-trained on a large news dataset.

We summarize our contributions as follows:

- We propose a new task of argumentative essay generation and create a large-scale and high-quality benchmark for this task.
- We establish a Transformer-based model with dual-decoder which generates argumentative essays in a plan-and-write manner, and further improve the model performance via pre-training.
- Using both automatic and human evaluations, we demonstrate that our proposed model can generate more coherent and persuasive argumentative essays with higher diversity and less repetition rate compared to several baselines.

2 Related Work

2.1 Argumentative Essay Analysis

The analysis of argumentative essays has been extensively studied in previous work since an early stage (Madnani et al., 2012; Beigman Klebanov and Flor, 2013). To comprehensively study the structure of argumentation in argumentative essays, Stab and Gurevych (2014, 2017) presented the Persuasive Essay dataset with the annotations of both argument components and argumentative relations. Based on this dataset, many subsequent researches are conducted to better parsing the argumentation

structure in argumentative essays (Persing and Ng, 2016; Eger et al., 2017; Potash et al., 2017; Kuribayashi et al., 2019; Bao et al., 2021).

These studies above are closely related to our work, since the analysis of the structure and quality of argumentative essays can support AEG by providing structured argument knowledge.

2.2 Argument Generation

Early work on argument generation involved a lot of hand-crafting features, such as constructing the argument knowledge base (Reed, 1999; Zukerman et al., 2000) or designing argumentation strategies (Reed et al., 1996; Carenini and Moore, 2000).

To frame existing argumentative text into new arguments, some work employs the argument retrieval (Levy et al., 2018; Stab et al., 2018) based methods to generate arguments (Sato et al., 2015; Hua and Wang, 2018; Wachsmuth et al., 2018), while others synthesize arguments by reframing existing claims or evidences (Yanase et al., 2015; Bilu and Slonim, 2016; Baff et al., 2019).

Recently, more attention has focused on end-to-end generation of arguments using neural models (Hua and Wang, 2018; Hidey and McKeown, 2019). Hua et al. (2019) presented a sequence-to-sequence framework enhanced by external knowledge for generating counter-arguments. Gretz et al. (2020) explored the use of a pipeline based on the pre-trained language model GPT-2 (Radford et al., 2019) to generate coherent claims. Schiller et al. (2021) developed a controllable argument generation model, which can control the topic, stance, and aspect of a generated argument. Alshomary et al. (2021a) proposed the belief-based claim generation task and leveraged conditional language models to generate arguments controlled by the prior beliefs of the audience. Khatib et al. (2021) proposed to control the generation of arguments with argumentation knowledge graphs.

However, current argument generation research is limited to generating individual and relatively short arguments, without consideration given to the generation of long and coherent argumentative essays containing multiple aspects of arguments.

2.3 Long-form text generation

Our work is also closely related to long-form text generation research, such as story generation (Fan et al., 2018; Yao et al., 2019; Guan et al., 2020; Xu et al., 2020), data-to-text generation (Puduppully et al., 2019; Hua et al., 2021; Hua and Wang, 2020; Dong et al., 2021), paragraph generation (Hua and Wang, 2019; Yu et al., 2021), and essay generation (Feng et al., 2018; Yang et al., 2019; Qiao et al., 2020; Liu et al., 2021).

Most of studies focus generating narrative texts or description texts, while we concentrate on generating argumentative essays, with more emphasis on the argumentativeness.

3 Dataset Creation

Our dataset is collected from Essay Forum,[2] an online community established by professional writers and editors to help users write, edit, and revise their essays. Specifically, we selected the essays and prompts of high-quality in the writing feedback section of Essay Forum, where users post their essays for revision suggestions in preparation for standardized English test like IELTS or TOEFL.[3] In addition, the essays in the writing feedback section have also been used in the researches on argument mining (Stab and Gurevych, 2014, 2017).

First, we collect all the post in the writing feedback section of Essay Forum. Then, to obtain the prompt-essay pairs and ensure the text quality, we conduct several pre-processing steps including:

- Separating the essay and the prompt in each post. For posts where the author does not mark the prompt in bold or italics, we filter them out and then process them manually;
- Filtering prompt-essay pairs with non-argumentative essays (like narrative essays, character description essays, and graphical analysis essays, etc.) by manually summarized rules (see Appendix B.1 for details.);
- Cleaning irrelevant text like special characters, user names, and expressions of thanks or greetings through rule-based deletion and manual processing (see Appendix B.2 for details.);
- Only keeping prompt-essay pairs whose essay contains less than 500 tokens (tokenized by the Stanford CoreNLP toolkit (Manning et al., 2014)) and 4 or 5 paragraphs. The reason for this procedure is that, in the writing feedback section of Essay Forum, essays that do not satisfy these aforementioned attributes are likely

[2] https://essayforum.com

[3] An example post in the Essay Forum can be found in Appendix A.

Dataset	Avg. Tokens	Avg. Sents
(Hua and Wang, 2018)	161.10	7.70
(Hua et al., 2019)	66.00	2.95
(Khatib et al., 2021)	81.89	3.85
ArgEssay (Ours)	327.35	14.41

Table 2: Comparison of our dataset with existing argument generation datasets. (Avg. Tokens)/(Avg. Sents) indicates the average number of tokens/sentences in the target generation text.

not in an argumentative writing style (Stab and Gurevych, 2014);

- Finally, manually reviewing each remaining prompt-essay pairs to filter obviously flawed essays and ensure all the essays are argumentative.

It is worth noting that the Essay Forum administrator will review and remove any posts that are considered to be libelous, racist, or otherwise inappropriate. Thus, the ethic of our dataset can be assured. Further, we also manually check the dataset to avoid ethical issues.

As for the data split, we want to minimize the overlap between the train set and the validation/test set in terms of prompts, otherwise it would be difficult to test the model's generalization ability on new prompts. Thus, we first extract keywords from the prompts based on TF-IDF (Salton and McGill, 1984) and measure the similarity of any two prompts as the Jaccard similarity between their keywords set. Then, when splitting the data, for any prompt in the validation/test set, we ensure that the similarity between it and each prompt in the train set does not exceed a threshold ϵ. After several rounds of manual verification, we set $\epsilon = 0.65$, as we observe that this threshold can reasonably separate the prompts with more than 70% of the validation/test prompts having a similarity of less than 0.30 to any training prompt.

The final dataset consists of $11,282$ prompt-essay pairs in English, in which $9,277/1,002/1003$ pairs are used for training/validation/testing, respectively. We compare our proposed dataset with existing argument generation datasets in Table 2. Our ArgEssay contains longer target text with richer content, which makes it more challenging. Also, most existing datasets are constructed from social media, while the essays in our dataset are written for the standardized English tests, which are more formal in terms of wording and structuring.

4 Methods

Our proposed AEG task can be formulated as follows: given a writing prompt $X = [x_1, x_2, \ldots, x_m]$, a relevant argumentative essay $Y^e = [y_1, y_2, \ldots, y_n]$ should be generated.

In order to generate diverse and content-rich essays, we propose a Transformer-based dual-decoder model with a plan-and-write strategy. In detail, our model first predicts a *planning* sequence Y^p, then it generates the argumentative essay Y^e under the guidance of the planning sequence through the planning attention. The planning strategy is commonly used in long-form text generation studies. Here, instead of using a standalone model for predicting the planning (Fan et al., 2019; Xu et al., 2020), we utilize a dual-decoder architecture to enable end-to-end training for generating the planning and the essay.

In the following, we will first introduce the method of constructing planning sequence Y^p for training and then describe our model in detail.

4.1 Construction of Planning

For flexibility, we do not strictly restrict the form of the planning, as long as it is natural language text. In this paper, we investigate two kinds of planning using on automatic methods, a **keyword-based planning** and a **relation-based planning**.

- 1) For the keyword-based (KW) planning, we use TF-IDF (Salton and McGill, 1984) score to determine important words as keywords. We calculate the TF-IDF based on the corpus and then select words with the top-l scores to construct the keyword-based planning $Y^p = k_1\#1|k_2\#2|\ldots|k_l\#l|$, where k_i is the i-th keyword, "#" and "i" are special tokens, and keywords are separated by "|".

- 2) Similarly, for the relation-based (Rel) planning, we firstly apply an off-the-shelf OpenIE (Angeli et al., 2015) to extract all the relational triplets in each essay and then random sample l triplets to construct the relation-based planning $Y^p = s_1\#r_1\#o_1\#1|\ldots|s_l\#r_l\#o_l\#l|$, where s_i, r_i and o_i are subject, relation and object of the i-th triplet.

Note that, we append "#i" after each keyword or each relational triplet to control the length of generated planning, which has been shown to prevent

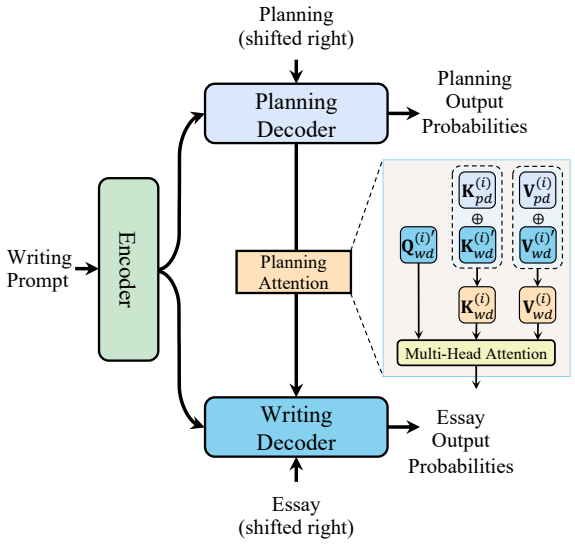

Figure 1: The architecture of our model.

the model from generating undesired excessive or insufficient keywords/triplets (Liao et al., 2019). Here, we refer to l as the planning length, and we set l to 10 in our main experiments. The impact of l is discussed in Section 6.5.

4.2 Dual-decoder Model

For essay generation task, we adopt the encoder-decoder architecture with a pre-trained BART backbone (Lewis et al., 2020) and extend it to a dual-decoder architecture. Figure 1 illustrates the overall architecture of the proposed dual-decoder. Overall, the proposed model consists of a shared encoder to encode the input writing prompt, a planning decoder to generate a planning sequence, and a writing decoder to write the argumentative essay.

Shared Encoder We use the same encoder of BART as the shared encoder in our model, whose output will be utilized by both decoders. Specifically, we feed X into the encoder:

$$\mathbf{H}^e = \text{Encoder}(X)$$

where $\mathbf{H}^e \in \mathbb{R}^{m \times d}$, and d is the hidden dimension.

Planning Decoder (PD) Based on the input prompt sequence, the planning decoder serves to predict the planning that contains important information of the essay. The generated planning can help plan the perspectives or aspects to be discussed in the essay before the formal writing, as well as enrich the wording and improve the diversity of the generated essay. Adopting the planning decoder allows to keep planning and writing process separate, with two decoders being responsible for each. The reason behind this design is that the distribution of the planning text and the essay text are significantly different, forcing one same decoder to handle both processes can decrease the performance.

Our planning decoder is based on the decoder of BART, whose decoding target text is Y^p:

$$\mathbf{h}_t^{pd} = \text{PD}(\mathbf{H}^e, Y^p_{<t})$$
$$\hat{Y}_t^p = \text{Softmax}(\mathbf{W}^{pd}\mathbf{h}_t^{pd} + \mathbf{b}^{pd})$$

where $\mathbf{h}_t^{pd} \in \mathbb{R}^d$ is the hidden representation of the t-th token in the generated logits $\hat{Y}_t^p$; $\mathbf{W}^{pd}$ and $\mathbf{b}^{pd}$ are learnable parameters.

Each Transformer layer of the BART decoder contains three sub-layers, i.e., a self multi-head attention layer, a cross multi-head attention layer and a feed-forward layer. For the self multi-head attention sub-layer of the j-th Transformer layer, we denote the keys and values matrix as $\mathbf{K}_{pd}^{(i)}$ and $\mathbf{V}_{pd}^{(i)} \in \mathbb{R}^{l \times d}$, which will be used to guide the writing decoder subsequently.

Writing Decoder (WD) The writing decoder can incorporates the generated planning and the input writing prompt to write an essay:

$$\mathbf{h}_t^{wd} = \text{WD}(\mathbf{H}^e, \mathbf{K}_{pd}, \mathbf{V}_{pd}, Y^e_{<t})$$
$$\hat{Y}_t^e = \text{Softmax}(\mathbf{W}^{wd}\mathbf{h}_t^{wd} + \mathbf{b}^{wd})$$

where $\mathbf{h}_t^{wd} \in \mathbb{R}^d$ is the hidden representation of the t-th token in Y^e; $\mathbf{K}_{pd}$ and $\mathbf{V}_{pd}$ are the keys and values of all the Transformer layers of PD; $\mathbf{W}^{wd}$ and $\mathbf{b}^{wd}$ are learnable parameters.

Here, we introduce a **planning attention (PA)** module that enables PD to guide WD. For each Transformer layer of the WD, we modify the self multi-head attention sub-layer to enable WD to attend all the tokens in the planning generated by PD when decoding each token of an essay. Specifically, when calculating the self multi-head attention in the i-th Transformer layer of WD, we use $\mathbf{Q}_{wd}^{(i)}$, $\mathbf{K}_{wd}^{(i)}$ and $\mathbf{V}_{wd}^{(i)}$ as the query, key and value:

$$\mathbf{Q}_{wd}^{(i)} = \mathbf{Q}_{wd}^{(i)'}$$
$$\mathbf{K}_{wd}^{(i)} = [\mathbf{K}_{pd}^{(i)} \oplus \mathbf{K}_{wd}^{(i)'}]$$
$$\mathbf{V}_{wd}^{(i)} = [\mathbf{V}_{pd}^{(i)} \oplus \mathbf{V}_{wd}^{(i)'}]$$

where $\mathbf{Q}_{wd}^{(i)'}, \mathbf{K}_{wd}^{(i)'}, \mathbf{V}_{wd}^{(i)'} \in \mathbb{R}^{n \times d}$ is the original query, key, value matrix of the BART decoder Transformer layer, and $\oplus$ denotes the matrix concatenation operation in the first dimension.

Training & Inference. During training, we use the negative log-likelihood loss:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_w$$
$$\mathcal{L}_p = -\sum_{t=1}^{l} \log P(Y_t^p | Y_{<t}^p, X)$$
$$\mathcal{L}_w = -\sum_{t=1}^{n} \log P(Y_t^e | Y_{<t}^e, Y^p, X)$$

where $\mathcal{L}_p$ and $\mathcal{L}_w$ are the loss functions for optimizing planning and writing, respectively.

During inference, we first generating the planning sequence and then writing the essay, both of which are performed in an autoregressive manner.

4.3 Pre-training

To better adapt the model to the plan-and-write paradigm, we explore to first pre-train our model on a large news dataset, then fine-tune it on our ArgEssay dataset. In detail, we employ CNN-DailyMail (Hermann et al., 2015) as the pre-training data, which is a large-scale news dataset commonly used for summarization. We treat the highlights as the prompts and the associated news articles as the essays. Regarding the planning sequences, the keywords/triplets are extracted from the news articles in the same way as described in Section 4.1.

5 Experimental Setups

5.1 Comparison Models

We build following baselines for comparison.

BART BART (Lewis et al., 2020) is a strong sequence-to-sequence baseline model for natural language generation, which is pre-trained on several denoising tasks. We fine-tune the pre-trained BART model on our proposed ArgEssay dataset without using any planning information.

BART-KW Following approaches of incorporating knowledge information with the arguments in previous work (Schiller et al., 2021), we conduct a BART-KW method by concatenating each planning before the essay as the overall target for prediction. That is BART-KW first predicts the keyword planning and then generates the essay. BART-KW is also fine-tuned from BART-base.

DD-KW For our dual-decoder (DD) models, we denote the dual-decoder model with keyword-based planning as DD-KW. Note that DD-KW is not pre-trained by news data but we use BART-base as the start point. Also, based on DD-KW, we implement following two models for further comparisons:

DD-KW *w/o planning-att* We make an ablation of planning attention module, that is we replace the planning attention for DD-KW with the normal attention, to investigate the effectiveness of using planning to explicitly guide essay generation. Note that this model differs from BART in that the planning can influence essay generation through the encoder during training.

DD-KW *w. pre-training* We apply the news pre-training on DD-KW (see Section 4.3).

BART-Rel and DD-Rel Similar for the methods using relation-based planning, we implement four models: **BART-Rel**, **DD-Rel**, **DD-Rel *w/o planning-att*** and **DD-Rel *w. pre-training***.

5.2 Implementation Details

For all models, we use the pre-trained BART-Base as the base model. Following previous work (Gretz et al., 2020; Xu et al., 2020; Khatib et al., 2021), for decoding at inference, we used a top-k sampling scheme with k = 40 and a temperature of 0.7. Our model is implemented in PyTorch (Paszke et al., 2019) and is trained on a NVIDIA Tesla V100 GPU. We restrict the generated text to be longer than 200 tokens. The AdamW optimizer (Kingma and Ba, 2015) is employed for parameter optimization with an initial learning rate of 3e-5.

5.3 Evaluation Metrics

Automatic Evaluation. We employ the following metrics for automatic evaluation. (1) **Distinct** measures the diversity of generated essays by computing the ratio of the distinct n-grams to all the generated n-grams (Li et al., 2016). (2) **Novelty** measures the difference between the generated essays and the training data. Specifically, following Yang et al. (2019) and Zhao et al. (2020), for each generated essay, we calculate its Jaccard similarity coefficient based on n-grams with every essay in the training set and choose the highest similarity as the novelty score. (3) **Repetition** measures the redundancy of the generated essay by computing the percentage of generated essays that contain at least one repeated n-gram (Shao et al., 2019). (4) **BLEU** (Papineni et al., 2002) computes the n-gram overlap between the generated texts and the reference texts. If the readability or fluency of the

Models	Diversity				Quality		
	Dist-3	**Dist-4**	**Nov-1**(⇓)	**Nov-2**(⇓)	**Rep-3**(⇓)	**Rep-4**(⇓)	**BLEU-4**
BART	46.68	70.43	26.73	9.45	19.04	3.09	6.85
BART-KW	48.95	72.18	26.67	9.31	17.24	2.89	6.74
DD-KW	50.07	72.72	†26.31	9.29	†16.87	2.55	6.81
w/o planning-att	47.13	70.76	26.78	9.43	18.74	†2.51	6.79
w. pre-training	**51.35**	**73.71**	**26.26**	†9.21	**16.75**	**2.39**	**6.94**
BART-Rel	47.45	71.39	27.41	9.48	21.14	3.29	6.72
DD-Rel	49.10	72.55	26.99	9.34	19.24	2.67	6.83
w/o planning-att	47.16	70.63	26.78	9.46	19.34	3.09	†6.93
w. pre-training	†51.11	†73.57	26.75	**9.20**	19.18	**2.39**	6.84

Table 3: Automatic evaluation results [%]. **Dist-n**, **Nov-n**, **Rep-n** and **BLEU-n** denote the distinct, novelty, repetition and BLEU based on n-gram. The best score is in bold. † indicates the second best result.

Models	Rel.	Coh.	Cont.
BART	3.27	2.83	3.09
BART-KW	3.31	2.71	3.31
DD-KW	3.60	2.83	3.42
w. pre-training	**3.63**	3.05	**3.49**
BART-Rel	3.27	2.78	3.29
DD-Rel	3.59	2.82	3.36
w. pre-training	3.60	**3.06**	3.43

Table 4: Human evaluation results. **Rel.**, **Coh.** and **Cont.** indicate relevance, coherence and content richness, respectively.

generated essay is poor, its BLEU score will be extremely low. Hence, we provide the BLEU score as a reference to assess the essay's quality.

Here, distinct and novelty are used for assessing diversity, while repetition and BLEU are used for assessing quality.

Human Evaluation. For a more comprehensive analysis, we conduct human evaluations that contain three aspects. (1) **Relevance** evaluates whether the entire content of the generated essay is semantically relevant to the given writing prompt, which is a basic requirement for a qualified argumentative essay. (2) **Coherence** indicates whether the generated essay is logically consistent and reasonable in terms of semantic and causal dependencies in the context, which is closely related to the persuasiveness of an argumentative essay. (3) **Content Richness** measures the amount of distinct relevant aspects covered in the generated essay, which is a significant characteristic of argumentative essays.

All three aspects are expected to be scored from 1 (worst) to 5 (best). We randomly sampled 50 writing prompts from the test set. Each annotation item contains the input writing prompt and the generated essays of different models. We assign 3 annotators for each item who are not aware of which model the generated essays come from.

6 Results and Analysis

6.1 Automatic Evaluation

Table 3 shows the automatic evaluation results. Compared to BART, our proposed DD-KW and DD-Rel achieve significantly better distinct scores and moderately better repetition and novelty scores. BART-KW and BART-Rel are worse in distinct, repetition, and novelty than DD-KW and DD-Rel, showing the effectiveness of the dual-decoder architecture. Also, removing the planning attention (*w/o planning-att*) decreases the distinct and repetition scores. Regarding the BLEU scores, DD-KW and DD-Rel perform similar to BART, indicating that the dual-decoder architecture does not degrade the readability and fluency of the generated essays. In addition, incorporating pre-training into our dual-decoder models can further boost the performance, showing that pre-training can enhance this plan-and-write generation paradigm. The average length of the essays generated by each models is around 290-300.

Overall, with the support of the dual-decoder architecture and the pre-training strategy, our model can generate more diverse and less repetitive essays at the same time maintaining good readability and fluency.

Models	Rec.	Rep.(⇓)	Inv.(⇓)	Rel.
BART-KW	18.06	6.45	-	77.40
DD-KW	19.41	1.80	-	82.00
w. pre-training	**23.95**	**1.01**	-	**84.80**
BART-Rel	14.81	-	1.76	72.20
DD-Rel	15.05	-	0.85	76.60
w. pre-training	**15.43**	-	**0.40**	**78.40**

Table 5: Planning quality evaluation [%]. **Rec.**, **Rep.**, **Inv.** and **Rel.** indicate recall, keyword repetition, invalidity and planning relevance, respectively.

Models	Appearance	Appropriateness
BART-KW	63.72	66.80
DD-KW	66.63	71.60
w/o planning-att	43.66	47.60
w. pre-training	**72.58**	**73.20**
BART-Rel	43.43	43.40
DD-Rel	51.31	52.40
w/o planning-att	19.01	37.20
w. pre-training	**52.99**	**57.40**

Table 6: Controllability evaluation [%].

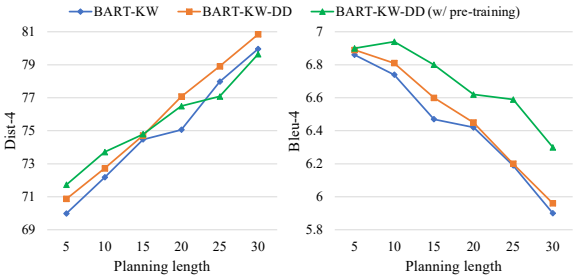

Figure 2: Impact of the length of planning.

6.2 Human Evaluation

The results of human evaluation are presented in Table 4. The average Fleiss' kappa is 0.42. Regarding relevance, BART, BART-KW, and BART-Rel perform poorly because of the topic drift problem, that is, the generated essay is barely relevant to the given topic (see case study in Appendix C for details). Compared to BART, all other models with planning achieve better content richness score, since the generated planning can provide more diverse aspects information and guide the models to write essays containing more examples or perspectives. Also, the pre-training strategy can bring significant improvement to coherence.

6.3 Planning Quality

We measure the quality of the generated plannings from the following aspects: (1) **Recall:** evaluates how many keywords/triplets in the oracle planning sequence are predicted. (2) **Keyword Repetition:** (only for keyword-based planning) measures how many keywords in the generated planning sequence are repeated at least once. (3) **Invalidity:** (only for relation-based planning) measures how many generated triplets is invalid, i.e., not in the form described in Section 4.1. (4) **Planning Relevance:** evaluates whether each predicted keyword/triplets is relevant to the prompt, and is obtained by manually analysis of 50 randomly selected samples.

As shown in Table 5, simply using a single decoder to generate the planning and the essay together (BART-KW and BART-Rel) causes the problem of high keyword repetition or high invalidity rate. In contrast, employing an individual planning decoder (DD-KW and DD-Rel) not only improves both the recall and the planning relevance, but also alleviates the keyword repetition or invalidity problem. Moreover, we can also observe that the planning quality can further be refined by pre-training our dual-decoder models.

6.4 Controllability Evaluation

To evaluate how well the generated essay can be controlled by the planning, we measure whether each keyword/triplet appear in the generated essay (**Appearance**). Also, we manually check 50 generated samples and determine whether the information enclosed by each keyword/triplet is appropriately used (**Appropriateness**). As shown in Table 6, BART-KW and BART-Rel achieve low appearance and appropriateness, while our dual-decoder models (DD-KW and DD-Rel) give significantly better results. With pre-training, around 73.20%/57.40% keywords/triplets are appropriately adopted by the writing decoder, showing a high controllability. Besides, removing the planning attention module (*w/o planning-att*) decreases both appearance and appropriateness dramatically.

6.5 Impact of the Planning Length

On top of the models with keyword-based planning, we further investigate the impact of the planning length l on the diversity (Dist-4) and accuracy (BLEU-4). As shown in Figure 2, for all models, as the planning length grows, the diversity increases, but the accuracy decreases. By manual review, we find that the readability of essays becomes extremely poor (low fluency and high repetition) when BLEU-4 is less than about 6.3. Thus, selecting a proper planning length is crucial for generating essays that are both diverse and readable.

Nevertheless, our pre-trained dual-decoder model (DD-KW *w. pre-training*) can not only achieve better diversity with an appropriate planning length, but also ensure better readability than baselines even under extreme conditions.

7 Conclusion

In this paper, we propose a challenging new task, AEG, to generate long-form and coherent argumentative essays. To tackle this task, we present a large-scale dataset and further devise a dual-decoder architecture based on the basis of BART, which can generate a planning and a planning-guided essay in an end-to-end fashion. The experimental results demonstrate the superiority of our model. For future work, we plan to draw on external knowledge to generate more diverse and informative argumentative essays.

Limitations

First, as discussed in Appendix C, there is still an undeniable gap between generated essays and human written essays in terms of logical coherence. In our method, we do not design mechanisms to ensure factual and causal logicality of the generated essays, which remains a great challenge. Hence, future work could consider improving the logical coherence of the generated essays by using external knowledge or causal inference techniques.

Second, although our dual-decoder architecture enables content planning and generates better essays, it also introduces some new parameters and computations. Future work could thus investigate more efficient methods with fewer model parameters.

Ethics Statement

Our dataset is collected from publicly available sources without any personal identity characteristics. When crawling data from the online platform "essayforum.com", we carefully read and follow the privacy policy, terms of use of this platform. According to the agreement of this platform, any content in it can be accessed and used with an indication of the source.

Since the administrators of the online platform we use will review and remove any posts that are considered to be libelous, racist, or otherwise inappropriate, the ethic of our dataset can be assured. We also manually double-check each sample in our dataset to confirm that no ethical issues exists.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China 62006062 and 62176076, Shenzhen Foundational Research Funding JCYJ20200109113441941, JCYJ20210324115614039, the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

References

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021a. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 224–233. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021b. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, Online. Association for Computational Linguistics.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 344–354. The Association for Computer 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, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 54–64. Association for Computational Linguistics.

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6354–6364. Association for Computational Linguistics.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013:i–15.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Giuseppe Carenini and Johanna D. Moore. 2000. A strategy for generating evaluative arguments. In *INLG 2000 - Proceedings of the First International Natural Language Generation Conference, June 12-16, 2000, Mitzpe Ramon, Israel*, pages 47–54. The Association for Computer 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Xiangyu Dong, Wenhao Yu, Chenguang Zhu, and Meng Jiang. 2021. Injecting entity types into entity-guided text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 734–741. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 11–22. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2650–2660. Association for Computational Linguistics.

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. 2018. Topic-to-essay generation with neural networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4078–4084. ijcai.org.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family - A study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 528–544. Association for Computational Linguistics.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Trans. Assoc. Comput. Linguistics*, 8:93–108.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6379–6393. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1756–1767. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2661–2672. Association for Computational Linguistics.

Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. 2021. DYPLOC: dynamic planning of content using mixed language models for text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6408–6423. Association for Computational Linguistics.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 219–230. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 591–602. 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 2020, Online, November 16-20, 2020*, pages 781–793. Association for Computational Linguistics.

Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4744–4754. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, and Kentaro Inui. 2019. An empirical study of span representations in argumentation structure parsing. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4691–4698. Association for Computational Linguistics.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2066–2081. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Yi Liao, Yasheng Wang, Qun Liu, and Xin Jiang. 2019. Gpt-based generation for classical chinese poetry. *CoRR*, abs/1907.00151.

Zhiyue Liu, Jiahai Wang, and Zhenghong Li. 2021. Topic-to-essay generation with comprehensive knowledge enhancement. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part V*, volume 12979 of *Lecture Notes in Computer Science*, pages 302–318. Springer.

Nitin Madnani, Michael Heilman, Joel R. Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 20–28. The Association for Computational Linguistics.

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 the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1384–1394. The Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1364–1373. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6908–6915. AAAI Press.

Lin Qiao, Jianhao Yan, Fandong Meng, Zhendong Yang, and Jie Zhou. 2020. A sentiment-controllable topic-to-essay generator with topic knowledge graph. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3336–3344. Association for Computational Linguistics.

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

Chris Reed. 1999. The role of saliency in generating natural language arguments. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 876–883. Morgan Kaufmann.

Chris Reed, Derek Long, and Maria Fox. 1996. An architecture for argumentative dialogue planning. In *International Conference on Formal and Applied Practical Reasoning*, pages 555–566. Springer.

Gerard Salton and Michael McGill. 1984. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, System Demonstrations*, pages 109–114. The Association for Computer Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 380–396. Association for Computational Linguistics.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3255–3266. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Demonstrations*, pages 21–25. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510. ACL.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Comput. Linguistics*, 43(3):619–659.

Stephen E. Toulmin. 2003. *The Uses of Argument*, 2 edition. Cambridge University Press.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 241–251. Association for Computational Linguistics.

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 2020, Online, November 16-20, 2020*, pages 2831–2845. Association for Computational Linguistics.

Toshihiko Yanase, Toshinori Miyoshi, Kohsuke Yanai, Misa Sato, Makoto Iwayama, Yoshiki Niwa, Paul Reisert, and Kentaro Inui. 2015. Learning sentence ordering for opinion generation of debate. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*, pages 94–103. The Association for Computational Linguistics.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2002–2012. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7378–7385. AAAI Press.

Wenhao Yu, Chenguang Zhu, Tong Zhao, Zhichun Guo, and Meng Jiang. 2021. Sentence-permuted paragraph generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5051–5062. Association for Computational Linguistics.

Liang Zhao, Jingjing Xu, Junyang Lin, Yichang Zhang, Hongxia Yang, and Xu Sun. 2020. Graph-based multi-hop reasoning for long text generation. *CoRR*, abs/2009.13282.

Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG 2000 - Proceedings of the First International Natural Language Generation Conference, June 12-16, 2000, Mitzpe Ramon, Israel*, pages 55–62. The Association for Computer Linguistics.

Appendices

A An Example Post from Essay Forum

An example post from the writing feed-back section of the Essay Forum platform is shown in Figure 3.

B Rules For Data Pre-process

B.1 Filtering Details

- Removing prompt-essay pairs which are from IELTS writing task 1 by checking for the presence of keywords like "bar", "chart", "diagram" and "task 1" in the prompts, since essays in these samples are graphical analysis essays without argumentativeness.
- Removing prompt-essay pairs about narrative, character description or letter by checking the keywords such as "describe", "describing", "letter", "narrative", "summary", etc.

B.2 Data Cleaning Details

- Deleting special characters like "=", "*", "#", "+", etc.
- Select out prompt-essay pairs that contain irrelevant text expressing gratitude, asking for help, greeting, or self-introduction by keywords like "please", "pls", "grammar", "hello", "feedback", "grammar", "comment", "my name", "my essay", "thank", "appreciated", etc. Then manually checking and deleting these irrelevant text.

C Case Study

Table 7 demonstrates several sample outputs from different models for the writing prompt about "multinational companie". We only show a snippet of each essay, which is taken from a similar location in the context.

We can see that BART and BART-KW show different degrees of topic drift problem, i.e., the generated text is less relevant to the given topic of "multinational corporations". In contrast, the models with dual decoders avoid this problem by better generating and utilizing the essay content plannings. Regarding the planning generation, BART-KW suffers from generating planning with repeated keywords. Also, as can be seen, the pre-trained dual-decoder models can better leverage the planning to guide the essay generation. For example, the generated essay

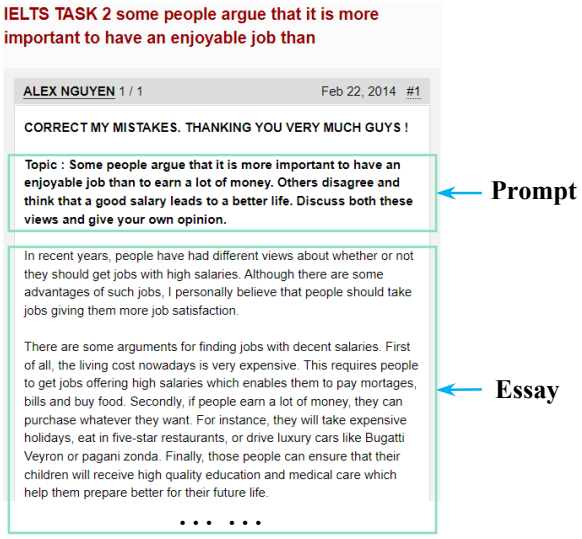

Figure 3: An example post from Essay Forum.

of (DD-KW *w. pre-training*) can coherently embody the keywords "employees", "coexist", "corporations", etc. By comparison, the DD-KW without pre-training performs slightly worse on this point.

However, it is undeniable that there is still a large gap between the performance of our models and that of humans, which we plan to investigate in future work. For example, the human generated essay in Table 7 is significantly better than the all the model generated essays in terms of logical coherence.

Writing Prompt: Nowadays, most large companies operate multinationally. To what extent those companies should have responsibility towards local communities in which they operate?
Human: *Keywords*: local # 1 | native # 2 | firms # 3 | royal # 4 | respect # 5 | rivers # 6 | operations # 7 | indigenous # 8 | international # 9 | ecosystem # 10 | *Triplets*: global operations # respect # national customs # 1 | firms # should discover # insights # 2 | operations # stay away # taboos # 3 | (. . .) *Essay*: (. . .) More importantly, global operations ought to respect the national customs where they set up. In other words, the firms should discover insights into the native cultures and deter themselves from offending indigenous people or what they respect. For example, when running business in Thailand, the organizations cannot advertise their products by using pictures of royal Kings as Thai citizens extremely respect royal family. Therefore, international operations should develop carefully and stay away from some taboos, which offend the inhabitants. (. . .)
BART: *Essay*: (. . .) Secondly, the public of the city can know about the local communities. They should be aware of the situations which they are dealing with. To elaborate, the local governments should make a better understanding of the cultures of the districts and they should also pay attention to the surrounding communities. Furthermore, they should be a part of the local companies' culture. (. . .)
BART-KW: *Keywords*: operate # 1 | companies # 2 | corporations # 3 | localities # 4 | local # 5 | multination # 6 | multi # 7 | multi # 8 | responsibility # 9 | response # 10 | *Essay*: (. . .) Additionally, the governments should also take a responsibility for the local companies. As the workers are being recruited more and more, the profits from these organizations will increase. For example, the factories can operate in different areas, in order for the workers to get a better life. Moreover, the workers will be more motivated to work more and have more opportunities to earn more income. (. . .)
DD-KW: *Keywords*: multinational # 1 | teamwork # 2 | scale # 3 | workers # 4 | operate # 5 | company # 6 | local # 7 | ethos # 8 | collaboration # 9 | businesses # 10 | *Essay*: (. . .) On the other hand, multinational companies have a large number of employees. Many of them spend most of their time in a different place and do not have the time to engage with the other employees. Moreover, The employees are boundlessly affected by the cultures, traditions, customs, and traditions. Therefore it is necessary for them to take a team effort to build a good relationship with the others and make sure that they have a good collaboration. (. . .)
DD-KW *w. pre-training*: *Keywords*: operate # 1 | profit # 2 | corporations # 3 | coexist # 4 | cities # 5 | multinational # 6 | local # 7 | responsibility # 8 | employees # 9 | development # 10 | *Essay*: (. . .) On the other hand, the main responsibility of the large multinational corporations is to coexist with the communities in the local areas, and they should share the costs and amenities of their business. In other words, they should be aware of the needs of their employees and the benefits of their job. Furthermore, they have to offer equal opportunities to all of their staffs. For instance, in Bangladesh, the government has made it mandatory for the companies to co-operate with the local community in order to facilitate the overall development of their businesses. (. . .)
BART-Rel: *Triplets*: factories # have responsibility # to # provide # 1 | companies # invest # money # 1 # 2 | local businesses # have # responsibility # 3 | (. . .) *Essay*: (. . .) Furthermore, it also can be argued if the local government has more responsibility for the well-being of their citizens. In some cases, they could not provide the basic needs for their citizens, such as food and clothing. If a company only has a little staff, they have to make sure their local residents have healthy food and a stable living environment. (. . .)
DD-Rel: *Triplets*: companies # are # able # 1 | their work # is with # their employee # 2 | many large enterprises # operate # multinationally # 3 | (. . .) *Essay*: (. . .) I think that a large company would have some responsibilities in the areas where they operate. Many multinational companies operate multilaterally, so they should have a responsibility towards the areas in which they operate in. In other words, they should be accountable for the way they operate from one part of the world to another. If these companies were able to build a business with their employees, they would have the ability to control the way that they operate, which would create a big impact on their business. (. . .)
DD-Rel *w. pre-training*: *Triplets*: companies # should have # their efforts # 1 | their efforts # is with # different initiatives # 2 | local people # is in # their local areas # 3 | most large firms # operate # multinationally # 4 | (. . .) *Essay*: (. . .) On the other hand, these multinationally large firms should have their own efforts regarding the local community. It is believed that their efforts can improve the situation of the community and people in their local areas. For instance, it could be better to invest in infrastructure that could improve the lives of the residents. Moreover, it allows them to start their efforts with different initiatives that would help them to increase their efficiency. (. . .)

Table 7: Case study.