Attend to the Right Context: A Plug-and-Play Module for Content-Controllable Summarization

Anonymous ACL submission

Abstract

Content-Controllable Summarization generates summaries focusing on the given controlling signals. We propose a plug-and-play module RelAttn to adapt any general summarizers to the content-controllable summarization task in zero- or few-shot settings. RelAttn first identifies the relevant content in the source documents, and then guides the model to attend to the appropriate context by directly steering the attention weight. We further propose an unsupervised online adaptive parameter searching algorithm to determine the degree of control in the zero-shot setting, while such parameters are learned in the few-shot setting. Experiments on three backbone models show that our method effectively improves all the summarizers, and outperforms both the prompting-based method and a widely used plug-and-play model.[1]

1 Introduction

Abstractive summarization is the task of generating a concise text containing the most relevant points for the given document. It is a widely explored task that has been greatly advanced by the use of large-scale generative language models trained on large corpora. The state-of-the-art abstractive summarizers (Zhang et al., 2020; Lewis et al., 2020; Xiao et al., 2021; He et al., 2022) achieve good performances regarding both salience and fluency. Following a long tradition in pre-neural query-based summarization (Dang, 2005; Rosner and Camilleri, 2008), researchers have just begun to investigate the neural controllable summarizers, which are expected to generate summaries that fulfill certain constraints, either on the format, e.g. the length (Liu et al., 2022) and style (Cao and Wang, 2021), or on the content (Dou et al., 2021; He et al., 2020), regarding specific entities, topics, and aspects. In this work, we focus on the content-controllable summarization task.

The task is essential in many practical settings. For example, users with specific information needs may be interested in different aspects or topics even for the same news article. While there are several large naturally annotated datasets for the task of generic summarization (like news with highlights and scientific papers with abstracts), no similar large corpora exists for content-controllable summarization. Furthermore, it is not financially feasible to annotate such a large corpus, which calls for zero- or few-shot methods. To take advantage of the general summarizers pre-trained on large-scale datasets, we propose a simple yet effective trainable plug-and-play module, RelAttn, which can effectively adapt the general summarization models to controllable summarization in both zero-shot and few-shot settings by steering attention weights.

Our proposal is inspired by findings in cognitive science, namely that when performing task-oriented reading comprehension tasks like question answering or summarization, humans use *selective attention* (Stevens and Bavelier, 2012; Dayan et al., 2000; Lavie et al., 2004), i.e., they focus on task-relevant information while suppressing any distractions. And intriguingly *selective attention* can be further trained to improve the reader's reasoning ability (Mccrudden et al., 2011).

As an essential component in the sequence-to-sequence transformer models, the Cross Attention, which guides the decoder to focus on a certain context of the input, plays a similar role as *selective attention*. For general summarizers, the attention is trained to focus more on the salient content. However, in the controllable summarization task, given the controlling signals, the selective attention should focus more on the relevant parts, not only the salient parts. Thus, to make the model focus on the updated task-relevant information, we propose to inject an adaptable Relevance Attention component into Cross Attention. We combine the Relevance and the Cross Attention with an adapt-

[1] The code will be made public once published.

able weight, determining how much controlling is needed. Inspired by Minimum Bayes Risk Decoding methods(MBRD, Suzgun et al. (2022)), we further propose an online adaptive hyper-parameter search algorithm (OS), which can be used to determine the degree of control for each single data example in fully unsupervised settings.

One close task explored recently is *guided summarization* (Dou et al., 2021; He et al., 2020; Narayan et al., 2021), with predicted or oracle keywords as the guidance. However, such guidance enhances the model ability to identify the salient content, therefore, improving the faithfulness of generated summaries, instead of making the model focus on the relevant content (Maddela et al., 2022).

We evaluate our method on two new annotated datasets (Maddela et al., 2022; Bahrainian et al., 2022), consisting of human-written summaries associated with different controlling signals. Experiments in both zero- and few-shot settings show that our module improves the performance of three summarizers (Lewis et al., 2020; Zhang et al., 2020; He et al., 2020) on both datasets, and our method outperforms or performs similarly to previous prompting-based methods and a plug-and-play method. Tellingly, more benefit is observed in the scenarios when higher degree of control is needed.

2 Related Work

2.1 Controlled Generation

Recent works on Controlled Text Generation either add a prefix/prompt at the beginning of the input (Keskar et al., 2019) or add a plug-and-play component to the large model (Dathathri et al., 2020; Krause et al., 2021; Yang and Klein, 2021).

In particular, PPLM (Plug-and-Play Language Model) proposed by Dathathri et al. (2020) combines an additional attribute model with a pre-trained unconditional language model to control the text generation. It uses gradient ascent on the LM's hidden activations to guide the generation of the next token to satisfy the control while maintaining fluency. Then it is extended by Krause et al. (2021), who employs a generative discriminator to improve the efficiency. Even more recently, Yang and Klein (2021) instead employs the attribute models to re-weight the output distribution of the LM considering an estimation on the controlling satisfaction of the partially generated text at each decoding step, archiving similar results. In this work, we compare our method with PPLM (Dathathri et al., 2020) on the datasets for the content controllable summarization task.

2.2 Controllable Summarization

Although generic summarization has received much-renewed attention in the deep-learning era (Zhang et al., 2020; Xiao et al., 2021; He et al., 2022; Chen et al., 2022), neural controlled summarization is still largely unexplored. In essence, there are mainly two kinds of controlling aspects: the format of generated summaries (e.g. length (Liu et al., 2022) or simplicity (Cao and Wang, 2021)) or the content of the generated summaries (e.g. regarding specific entities and topics they should focus on, queries or questions they should answer). Recently, Zhang et al. (2022) propose a benchmark for controlled summarization with multiple attributes, and GPT-like models have been applied with limited success to the aspect-based summarization task (Goyal et al., 2022). Closely relevant to content-controllable summarization, most recent works focus on the *guided-summarization task*. Specifically, one extracts the oracle entities or keyphrases from the ground-truth summaries, which are then used to guide the generation of the summaries. For example, Dou et al. (2021) uses an additional encoder to encode the guidance signals, with partial parameters shared with the original document encoder. In another line of research, He et al. (2020) propose a pre-training strategy prepending the source documents with the oracle keywords as a prompt, and Narayan et al. (2021) train the model to first predict an entity chain before predicting the final summary. However, different from the content-controllable summarization, the injection of those keywords enhances the model to find the most important content which improves the models' faithfulness, rather than making the model focus on relevant content in the context (Maddela et al., 2022).

Recently, Bahrainian et al. (2022) have collected a human-annotated topic-focused summarization dataset *NEWTS*. It was created by sampling from the CNNDM dataset using the topic information as the controlling signal. Similarly, Maddela et al. (2022) have introduced a human-annotated dataset *EntSUM* for entity-centric summarization. We evaluate our proposed method on these two datasets.

3 RelAttn: A Plug-and-Play Module

The content-controllable summarization task can be defined as follows: given a document with n

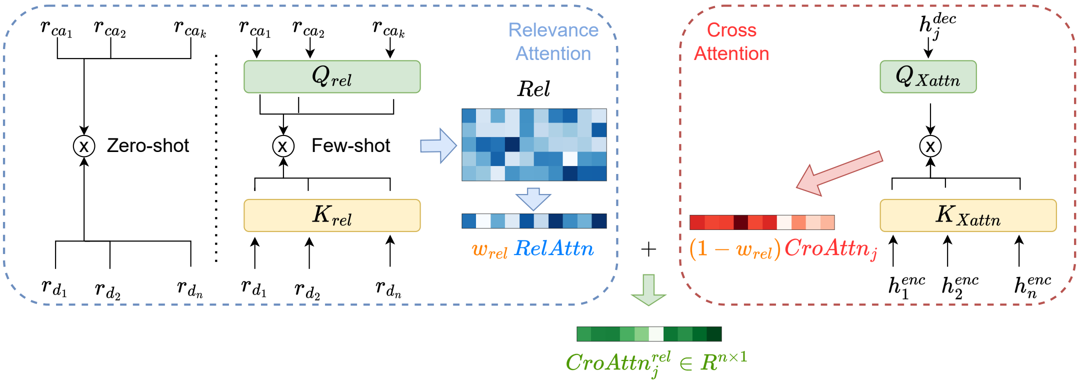

r_{ca_1} r_{ca_2} r_{ca_k}
Zero-shot
r_{d_1} r_{d_2} r_{d_n}
r_{ca_1} r_{ca_2} r_{ca_k}
Q_{rel}
Few-shot
K_{rel}
r_{d_1} r_{d_2} r_{d_n}
Relevance Attention
Rel
w_{rel} $RelAttn$
+
Cross Attention
h_j^{dec}
Q_{Xattn}
K_{Xattn}
$(1 - w_{rel})CroAttn_j$
h_1^{enc} h_2^{enc} h_n^{enc}
$CroAttn_j^{rel} \in R^{n \times 1}$

Figure 1: The illustration of the RelAttn module. The Relevance Attention is injected into the Cross Attentions with w_{rel}, indicating the degree of control. In the few-shot settings, Q_{rel} and K_{rel} are trainable linear projections.

tokens, $D = \{d_1, d_2, ...d_n\}$, and controlling aspects $CA = \{ca_1, ca_2, ...ca_k\}$, which can be a sequence of words, entities or key-phrases indicating the content of interest, the objective is to generate a summary that focuses on the controlling aspects.

Our main idea is to inject the Relevance Attention, indicating how important each input token is, into the Cross Attention of the transformer-based summarizers. The objective is to guide the model to 'pay more attention' to content relevant to the controlling aspects.

3.1 Cross Attention

There are three kinds of attention in a sequence-to-sequence transformer model - Encoder Attention, Decoder Attention, and Cross Attention. The Cross Attention, in particular, is used by the decoder to refer to the input sequence. As shown in Figure 1(right), in each cross attention head with model dimension dim, suppose $h_j^{dec} \in \mathbb{R}^{1 \times dim}$ is the output of the previous layer of the decoder at the decoding step j, and $h^{enc} \in \mathbb{R}^{n \times dim}$ is the output of the last layer in the encoder, then the cross attention vector is computed as:

$$CroAttn_j = Softmax(Q(h_j^{dec}) \cdot K(h^{enc})^T)$$

where $Q(\cdot)$ and $K(\cdot)$ are the query and key vectors respectively, and $CroAttn_j \in \mathbb{R}^{1 \times n}$ indicates the attention over the input tokens, with higher attention values on input tokens influencing more the generation of the current token.

3.2 Relevance Attention

To adapt the models to the controllable summarization task, we propose Relevance Attention to identify the *relevant content* in the input. And then such attention is injected into Cross Attention in all the heads, making the model focus on the content.

Specifically, as shown in Figure 1 (left), we denote the representation of each controlling aspect token ca_j and input token d_i as r_{ca_j} and r_{d_i}, respectively, and we directly use the embeddings of each token as the representation. Then the controlling aspect with k tokens and source input with n tokens can be represented as $r_{ca} \in \mathbb{R}^{k \times dim}$ and $r_d \in \mathbb{R}^{n \times dim}$, respectively. The relevance attention $RelAttn$ is computed either based on the dot product of the representations in the zero-shot settings or that of the key and query vectors learned in few-shot settings, i.e.

$$\text{(Zero-shot)} Rel = r_{ca} \cdot r_d^T$$

$$\text{(Few-shot)} Rel = Q_{rel}(r_{ca}) \cdot K_{rel}(r_d)^T$$

$$RelAttn = Softmax(\sum_k Rel)$$

, where $Q_{rel}(\cdot)$ and $K_{rel}(\cdot)$ are the query and key vectors in the proposed RelAttn component. The resulting $Rel \in \mathbb{R}^{k \times n}$ represents the relevance of every source document token with every controlling aspect token, thus $\sum_1^k Rel \in \mathbb{R}^{1 \times n}$ represents the overall relevance score of the source document tokens. In summary, $RelAttn_i$ measures how relevant the i-th input token is with the controlling aspects, and the model is encouraged to focus more on the relevant content.

As shown in Figure 1 (middle bottom), the relevance attention matrix is then combined with the original cross attention matrices of every attention head with a relevance weight w_{rel}, representing how much controlling is conducted, i.e.

$$CroAttn_j^{rel} = w_{rel} RelAttn + (1 - w_{rel}) CroAttn$$

$$output_j = CroAttn_j^{rel} \cdot V(h^{enc})$$

Dataset	# Examples	# Src Doc	#Summ/Doc	Len(doc)	Len(summ)	Ctrl Asp.
NEWTS	4800/1200	2388/574*	2	539	67	Topic Info
EntSUM	734/1994	164/481	1-18	861	95	Entity

Table 1: The statistics of both datasets.

, where $V(\cdot)$ represents the value vector in the Cross Attention, and $CroAttn_j \in \mathbb{R}^{1\times n}$ is the original cross attention over all the source input tokens at decoding step j. Thus, the final output is computed using the updated cross attention $CroAttn^{rel}_j$ with the proposed Relevance Attention included.

Gaussian Smoothing: To enhance the model's awareness of the context surrounding relevant words, we introduce Gaussian smoothing to the Relevance Attention module. This helps prevent attention vectors from becoming overly focused on highly related words, such as those identical to the controlling aspects, particularly in zero-shot settings. This approach aligns with recent research on attention weight smoothing (Lohrenz et al., 2022; Maekaku et al., 2022).

3.3 Determine the Degree of Control

In essence, w_{rel} is a hyperparameter indicating how much influence the controlling aspects are supposed to have on the model. In this section, we introduce practical ways to determine the value of w_{rel} in both zero-shot and few-shot settings.

3.3.1 Zero-shot: Online Adaptive Parameter Selection

In previous works, hyper-parameters like w_{rel} are usually set according to the performance on a small validation set (Zheng and Lapata, 2019), but this is not feasible for zero-shot. Also, different degrees of control may be needed, which requires different w_{rel} for each data example. We, therefore, propose an online adaptive parameter selection algorithm (OS) to select w_{rel} for each single data example, inspired by the recent works on Minimum Bayes Risk Decoding (MBRD, Suzgun et al. (2022)).

The key idea is to find the 'central' summary over a candidate set containing the summaries generated by the model with different w_{rel}.[2] Specifically, for each data example i, we generate n candidate summaries with different w_{rel} to form a candidate pool $C^i_{\{w_{rel}\}} = \{S^i_{w_{rel}}\}$. And then for each candidate, we compute an alignment score with all the other candidates. We finally pick the w_{rel} with the summary having the highest alignment score with other candidates, i.e.

$$\bar{w}^i_{rel} = \underset{w_{rel}}{argmax} \frac{1}{|C^i_{\{w_{rel}\}}|} \sum_{y\in\{w_{rel}\}} a(S^i_y, S^i_{w_{rel}})$$

In this way, the chosen value for w_{rel} is expected to have a more balanced influence on the model for the target data example.

3.3.2 Few-shot

In the few-shot setting, we make the model learn different w_{rel} for each head. Specifically, in each head, we apply a linear layer mapping the concatenation of the representations of source documents and controlling signals to a real number, followed by a sigmoid activation function to convert the number to probability space.

4 Datasets and Baselines

4.1 Datasets

We use two datasets for the experiments, controlling the content of generated summaries by topic information and entities. The detailed statistics of the datasets can be found in Table.1.[3]

NEWTS (Bahrainian et al., 2022) is a topic-oriented summarization dataset. For each article, there are two human-written summaries regarding two different topics, with the topics represented by a sequence of topic words and phrases, along with a topic sentence. The topic words and phrases are selected based on their high frequency within each topic, but they may not necessarily appear in both the source article and the human-written summaries. In our experiments, we focus solely on using the topic words for simplicity.

EntSUM (Maddela et al., 2022) is an entity-centric summarization dataset, where each example contains a document, an entity, and a human-written summary regarding the entity. The given entity is not necessarily central in the source document. The original dataset is test-only, for hyper-parameter search and few-shot experiments, we randomly split the dataset into train/test sets.

[2] We add some heuristic constraints to ensure the summaries in the candidate set are of reasonable quality. The details can be found in Section 5.1

[3] There are some duplicates in the NEWTS dataset, so the number of unique documents is slightly different from the expected number.

More details of the datasets are in Appendix A.

4.2 Models

To evaluate the effectiveness of our method, we use three pre-trained models as the backbone model, including a pre-trained model for general summarization (PEGASUS), a model fine-tuned for the general summarization task (BART-CNNDM, or in short BART[4]), and a model trained for conditional summarization task (CTRLsum).

In previous studies (Bahrainian et al., 2022), incorporating keywords as prompts has proven to be a successful approach for adapting a general summarizer to controlled summarization tasks, despite its simplicity. Therefore, we employ this efficient prompting-based method as a baseline in our experiments (-CA+Doc).

Besides, we also compare our method with the three backbone models, when they are extended with a widely used plug-and-play module (PPLM, Dathathri et al. (2020). The details of the models can be found in Appendix B.

5 Experiments and Analysis

5.1 Settings and Implementation

For all the pre-trained models (PEGASUS, BART-CNNDM, and CTRLsum) used in this paper, we directly use the publicly available checkpoints[5].

In the zero-shot setting, the relevance weight w_{rel} is selected in three ways, as listed below, from a candidate set, which contains all the 30 numbers within the range $[0.01, 0.30]$ with a step size 0.01. And we also add several heuristic constraints to ensure the summaries generated by w_{rel} in the candidate sets are reasonable: we remove all the candidate summaries with too many repeated words and the summaries distracted too much [6], and we also remove all the duplicated summaries. [7]

RelAttn(Val): w_{rel} is a fixed number shared for the whole dataset, set based on the performances on a randomly sampled validation set with size 100.

RelAttn(OS): w_{rel} is determined for each single data example using the online selection algorithm introduced in Section 3.3.1. We use ROUGE-1 as the alignment score in the OS algorithm.

RelAttn(OS-oracle): w_{rel} is set to be the number with which the summaries have the best ROUGE score with the ground-truth summaries. As such, it shows the *empirical* upper bound of the RelAttn model in the zero-shot setting.[8] More detailed settings are discussed in Appendix C.

5.2 Zero-Shot Results

We evaluate all models on the two datasets in the zero-shot setting using ROUGE scores (Lin, 2004) and BERTScore (Zhang* et al., 2020) as evaluation metrics. The results in Table 2 are organized in separate blocks for each of the three backbone models.

The scores between the general and controlled reference summaries, as shown in the first row, are notably low for both datasets. This suggests that the controlled summaries differ significantly from the general summaries.

RelAttn generally helps. With the proposed RelAttn module (-Doc v.s. +RelAttn), the performances of all three models on both datasets improve except for the CTRLsum on the NEWTS dataset, and most of the improvements are significant. In essence, the RelAttn module is more effective for all three models on the EntSUM dataset, and the reason might be that as the dataset is more extractive, paying attention to the correct context brings a more obvious improvement.

Simply adding the controlling aspect as a prompt provides limited help. Prompting-based method (-CA+DOC v.s. -DOC) often improves the model within a small margin, but sometimes even hurts the performance.

Pre-training for guided summarization does not always lead to a performance gain on the controllable summarization task. Overall, CTLR-Sum does not dominate the other two models, which shows that the model pre-trained for the guided summarization task may not be directly applied to the controllable summarization task.

Unsupervised OS algorithm for flexible w_{rel} outperforms fixed number selection. Except for the PEGASUS model on the NEWTS dataset, using the unsupervised OS algorithm to select w_{rel} for each data example yields better performance than using a fixed number based on a small validation

[4] All 'BART' mentioned refer to 'BART-CNNDM'

[5] https://huggingface.co/models

[6] By practice, we remove the summaries with unique word ratios less than 0.6 and with word overlap less than 0.2 with the generated summary without control.

[7] We set the upper bound to be 0.30 for computational efficiency, and by observation, for most of the data examples, the real upper bound for the filtered candidate set is below 0.30.

[8] Since we only compute it for 30 values, an even higher bound could be possible.

Model	NEWTS R-1	 R-2	 R-LSum	 BERTScore	EntSUM R-1	 R-2	 R-LSum	 BERTScore
General Summ	28.82	7.62	23.55	84.99	27.51	9.79	18.26	84.07
PEGASUS-Doc	33.21	11.41	29.06	**85.55**	36.44	18.75	32.76	85.72
PEGASUS-CA+Doc	**33.58**	**11.45**	29.02	85.20	36.99	19.08	33.19	85.69
PEGASUS+PPLM(Val)	32.11	10.88	28.08	85.32	34.90	17.10	31.17	85.29
PEGASUS+RelAttn(Val)	**33.37**	**11.50**	**29.28**	**85.59**	**37.49**†	**19.84**†	**33.84**†	**85.92**
PEGASUS+RelAttn(OS)	33.23	11.33	**29.20**	85.43	**39.02**†	**21.43**†	**35.36**†	**86.14**
PEGASUS+RelAttn(OS-oracle)	36.87	13.53	32.35	85.97	44.90	27.33	41.14	87.10
BART-Doc	33.09	10.67	28.98	**86.01**	30.50	12.10	27.33	85.16
BART-CA+Doc	32.90	11.01	29.48	85.98	30.73	12.46	27.59	85.17
BART+PPLM(Val)	31.61	10.30	28.67	85.70	29.22	10.56	26.25	84.67
BART+RelAttn(Val)	**33.10**	**11.11**	**29.70**	**86.01**	**33.47**†	**16.07**†	**30.55**†	**85.58**
BART+RelAttn(OS)	**33.34**†	**11.19**	**30.09**†	**86.06**	**34.42**†	**16.23**†	**31.21**†	**85.88**
BART+RelAttn(OS-oracle)	39.48	14.95	35.58	86.81	44.83	27.76	41.57	87.54
CTRLsum*	**32.57**	**9.58**	**29.06**	**85.52**	35.97	**18.56**	32.69	85.96
CTRLsum+PPLM(Val)	30.32	8.31	27.23	85.26	32.40	14.33	29.18	85.10
CTRLsum+RelAttn(Val)	32.12	9.17	28.45	**85.51**	**36.14**	18.40	**32.82**	**86.03**
CTRLsum+RelAttn(OS)	**32.38**	**9.53**	**28.91**	85.40	**37.24**†	**19.27**†	**33.71**†	**86.20**
CTRLsum+RelAttn(OS-oracle)	37.54	12.01	33.17	85.92	45.25	27.05	41.34	87.47

Table 2: Results on both datasets in zero-shot settings with three different backbone models. The top-2 performers (except the oracle ones) are bold. † indicates that the model is significantly better than the baseline models with $p < 0.05$ by bootstrap significant test (Graham et al., 2014).

Model	NEWTS	EntSUM
PEGASUS-Doc	26.91	21.39
PEGASUS-CA+Doc	**32.20**	24.41
PEGASUS+RelAttn(OS)	28.79	**25.78**
BART-Doc	26.29	20.42
BART-CA+Doc	26.60	20.70
BART+RelAttn(OS)	**28.90**	**25.95**
CTRLsum	42.02	30.52
CTRLsum+RelAttn(OS)	**43.56**	**35.72**

Table 3: SimCSE between the generated summary and the controlling aspects.

set. This indicates that different data examples benefit from varying degrees of control, which can be effectively determined by the OS algorithm.

The performances still have a large room to improve with a better selection algorithm for w_{rel}, as there is still a gap between the performance of RelAttn(OS) and RelAttn(OS-Oracle) - last row of each block.

RelAttn enhances similarity between generated summaries and controlling aspects. We assess the relevance of the generated summaries to the controlling aspects by computing the SimCSE score (Gao et al., 2021) between them, as shown in Table 3. The RelAttn module significantly improves the similarity between the generated summaries and the controlling aspects in most cases, except for its application to PEGASUS on the NEWTS dataset. This deviation can be attributed to the tendency of the pre-trained PEGASUS model to directly copy sentences from the source document, where the initial sentences are frequently copied. Consequently, the prompt (controlling aspects) is included in the summary, resulting in higher similarity with the controlling aspects.

5.3 Analysis on the Degree of Control

We further investigate the effectiveness of our method across varying levels of required control. To measure the expected degree of control, we utilize the similarity (ROUGE) between gold-standard controlled summaries and general summaries. Higher control is needed when the controlled summaries differ more from the general summaries. To analyze the performances of our method, we divide the data examples into bins based on the ROUGE-1 scores.

Figure 2 (left) shows the number of instances where RelAttn (OS) and the prompting-based method outperform each other for different levels of control. RelAttn tends to perform better when greater control is required, particularly when the ROUGE-1 scores between general and controlled summaries fall within the range of $[10, 40]$.

Next, we explore the influence of the parameter w_{rel} on the degrees of control (Figure 2, right). The expected trend is for higher values of w_{rel} to be selected for examples requiring more control. While the oracle w_{rel} (by OS-Oracle) aligns with this trend for both datasets, the average values of w_{rel} selected by the OS algorithm remain similar across all bins. One possible explanation is that our algorithm selects the proper w_{rel} using the "central" summary from the candidate set, which often coincides with the central value. However, future work can explore improving the unsupervised algorithm by considering the distance between salient and relevant content.

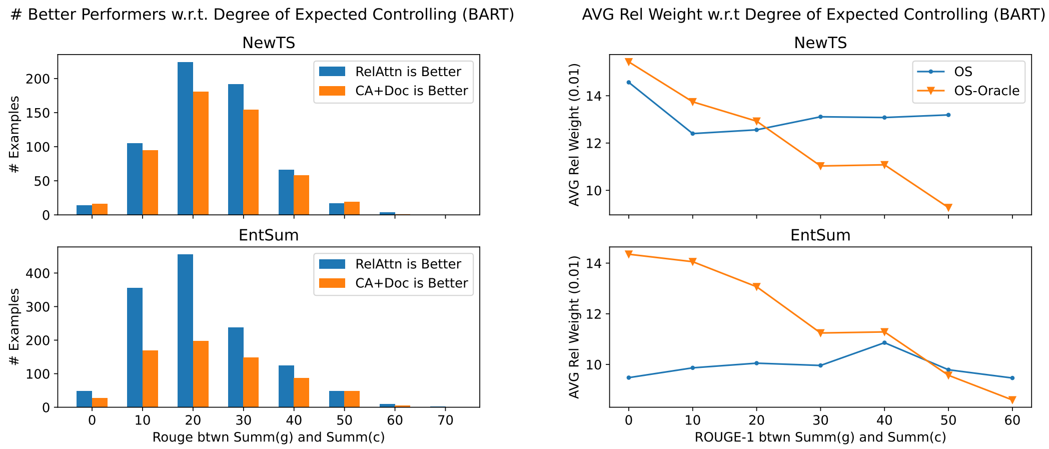


Figure 2: Analysis on the degree of expected control (BART). Left: The number of better performers among CA+Doc and RelAttn(OS). Right: The avg w_{rel} of OS and -Oracle.

Model	NEWTS R-1	R-2	R-L	EntSUM R-1	R-2	R-L
Pgeasus-CA+Doc	**33.63**†	**11.42**	**29.06**	37.12	19.31	33.30
PEGASUS+RelAttn (Ours)	33.19	11.32	29.03	**37.66**†	**20.25**†	**34.03**†
BART-CA+Doc	33.65	11.26	30.09	39.45	23.32	36.23
BART+RelAttn (Ours)	**34.36**†	**11.81**†	**30.67**†	**40.27**†	**24.63**†	**37.03**†
CTRLsum	33.54	11.01	29.92	51.57	39.83	49.00
CTRLsum+RelAttn (Ours)	**33.88**†	**11.07**	**30.18**†	**52.42**†	**40.65**†	**49.82**†

Table 4: Results on both datasets in few-shot settings with three backbone models. We train all the models on 10 data examples for 5 times, sampled with the same set of random seeds, and we show the mean ROUGE scores.

5.4 Few Shot Results

We train the prompting-based models and the RelAttn injected models using all three backbone models on both datasets. The training is conducted with 10 data examples, repeated 5 times. The average ROUGE scores are reported in Table 4. In most of the experiments, our method shows significant improvements over the prompting-based method. Comparing the models with the zero-shot results, CTRLsum makes the largest improvements, and PEGASUS makes the least. The reason might be that the guided summarizer is trained in the way closest to controllable summarization task, therefore even a small amount of in-domain data is sufficent to adequately adapt the model to the task.

5.5 Qualitative Analysis

Table 5 presents the results of a real example in the zero-shot setting from the EntSUM dataset. The article discusses crimes occurring in a computer chat room, with the given entity being one of the accused suspects (James Buckley). The general summarizer (BART-Doc) generates a summary that does not mention the entity specifically, but rather covers the general set of crimes. In contrast, with the RelAttn module, the model considers the entity, resulting in summaries that focus on the relevant content. However, if the value of w_{rel} is excessively large, the generated summary may consist of repetitive words instead of fluent text, highlighting the importance of selecting an appropriate w_{rel}. The trained guided-summarization model (CTRLsum) generates a summary that contains only one relevant sentence, while the rest focuses on the general content. Overall, the inclusion of the RelAttn module appears to enhance the focus on the given entity in the generated summary.

5.6 Human Evaluation

We conduct a human evaluation for zero-shot summaries with Amazon Mechanical Turk.[9]

We randomly selected 40 examples (20 from each dataset) that included two system summaries generated by BART-CA+DOC, BART-RelAttn(OS), and the reference summaries. Three annotations were collected for each example. Annotators were shown the original documents, controlling aspects, and the three summaries in a *random* order. They were asked to score each summary on Relatedness, Factualness, Grammaticality, Co-

[9] More details can be found in Section E.

Model Type	Summary
Ground-truth	The reporter, James Buckley, 50, of Brookhaven, who has worked for more than 20 years at WALK-FM (97.5), was arraigned on Tuesday morning in United States District Court in Central Islip. Allen Bode, an assistant United States attorney, said in a telephone interview that Mr. Buckley was caught trading pornography, including some hard-core material involving toddlers, with undercover agents in Germany and the United States. Mr. Buckley faces up to 19 years in federal prison if convicted.
BART-Doc $(w_{rel} = 0)$	24 men accused of soliciting sex from detectives posing as children in computer chat rooms. In a separate case, a traffic and weather reporter for a Long Island radio station was indicted on federal charges of possessing child pornography. All were charged with attempting to disseminate indecent material to a minor, a felony with a maximum penalty of four years in prison. ...
BART+RelAttn (OS-Oracle) $(w_{rel} = 0.12)$	Buckley faces up to 19 years in federal prison if convicted.Buckley has worked for more than 20 years at WALK-FM (97.5) Buckley's court-appointed defender, Randi Chavis, said in a telephone interview that he pleaded not guilty. Buckley was caught trading pornography, including some hard-core material involving toddlers, with undercover agents in Germany and the United States.Buckley's attorney declined to comment.
BART+RelAttn (OS) $(w_{rel} = 0.15)$	Buckley faces up to ...*(same as above)* declined to comment.Buckley is one of 11 men arrested after sending sexually explicit messages.
BART+RelAttn $(w_{rel} = 0.30)$	Buckley Buckley, Buckley Buckley Buckley is Buckley Buckley's Buckley Buckley.Buckley Buckley faces Buckley Buckley was ...
CTRLsum	James Buckley, 50, of Brookhaven, faces up to 19 years in federal prison if convicted. Police say they created online profiles and screen names for fictitious boys and girls. ... autoimmune autoimmune autoimmune disease. autoimmune disease, autoimmune disease of the mind, immune system, and body.
CTRLsum+RelAttn (OS) $(w_{rel} = 0.05)$	James Buckley, 50, of Brookhaven, faces up to 19 years in federal prison if convicted. Buckley was caught trading pornography, including hard-core material involving toddlers, with undercover agents in Germany and the U.S. Police say they created online profiles and screen names for fictitious boys and girls, ages 12 to 14, and visited chat rooms operated by AOL, Yahoo and other Internet service providers. ...

Table 5: An example from the EntSUM dataset (*doc_id: 1739833*) with the entity **James Buckley**, and the full version can be found in Table 8 in the Appendix.

Model	NEWTS					EntSUM				
	Rel	Fact	Gram	Coh	Overall	Rel	Fact	Gram	Coh	Overall
Reference	5.03	5.12	5.30	4.93	4.68	5.57	6.27	5.12	5.03	4.97
BART-CA+DOC	4.85	5.87	5.52	5.37	5.08	3.67	6.13	5.78	5.53	4.13
BART+RelAttn(OS)	5.18	5.87	5.68	5.55	5.32	4.78	5.97	5.30	5.22	4.85

Table 6: Results of human evaluation on both datasets, each with 20 examples. Annotators score each summary regarding Relatedness, Factualness, Grammaticality, Coherence and Overall quality, all the scores range within 1-7.

herence[10], and Overall quality using a scale of 1-7, where 1 indicated very poor and 7 indicated very good for each aspect. Results are shown in Table 6.

Our method is consistently better than the baseline model (BART-CA+DOC) on both datasets regarding Relativeness, and the Overall quality, while having a similar performance regarding Factualness, Grammaticality and Coherence. This demonstrates that our method effectively enhances the relatedness of the generated summaries without compromising fluency in the zero-shot setting. It is worth noting that both system summaries perform better than the reference summary in NEWTS, which could be attributed to variations in evaluation criteria applied by the reference summary annotators and the scoring annotators in assessing what constitutes a "good" summary.

6 Conclusion and Future Work

In this work, we propose a plug-and-play module, RelAttn, to adapt general summarization models to the content-controllable summarization task in zero/few-shot settings, which can be applied to any transformer-based sequence-to-sequence model. In essence, RelAttn helps the model focus on the content more relevant to the given controlling aspects by steering the attention weights in the cross attention. An online adaptive parameter selection algorithm is proposed in the zero-shot setting to estimate the degree of control, and such module is also trainable in the few-shot settings. We conduct experiments on two datasets with three summarization models as backbone models. The results show that RelAttn effectively adapts the general summarizers to content-controllable summarization in both zero- and few-shot settings, and the module is shown to be more helpful when more control is needed in the zero-shot settings.

Given the recent success of decoder-only GPT-like models (OpenAI, 2023) in many NLP tasks, we plan to study how to combine our Relevance Attention component with a decoder-only self-attention, in order to contribute to addressing their still existing limitations for aspect-based summarization, originally identified in Goyal et al. (2022).

[10] No specific methods were employed to enhance the factualness, grammaticality, and coherence of the summaries. Consequently, no performance improvement is expected in these aspects.

Limitations

We propose a plug-and-play module for adapting the general summarizers to the content-controllable summarization task in zero- or few-shot settings. Despite its effectiveness on the relatedness of the generated summaries toward the given controlling aspects, similar to most other plug-and-play modules, it may occasionally generate jerky text in the zero-shot settings. Though our proposal shows a competitive performance on fluency in the human evaluation, further research is still needed to improve the fluency of the generated summaries by the plug-and-play modules.

Ethics Concern

Similar to existing state-of-the-art generation models, there is no guarantee that our model will always generate factual content. Therefore, caution should be exercised when deploying the model in practical settings. Factuality is an open problem in existing generation models.

References

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Shuyang Cao and Lu Wang. 2021. Inference time style control for summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5942–5953, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. Unisumm: Unified few-shot summarization with multi-task pre-training and prefix-tuning.

Hoa Trang Dang. 2005. Overview of duc 2005.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Peter Dayan, Sham Kakade, and P Read Montague. 2000. Learning and selective attention. *Nature neuroscience*, 3(11):1218–1223.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Fatema Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *CoRR*, abs/2012.04281.

Pengcheng He, Baolin Peng, Liyang Lu, Song Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, et al. 2022. Z-code++: A pre-trained language model optimized for abstractive summarization. *arXiv preprint arXiv:2208.09770*.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

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*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nilli Lavie, Aleksandra Hirst, Jan W De Fockert, and Essi Viding. 2004. Load theory of selective attention and cognitive control. *Journal of experimental psychology: General*, 133(3):339.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2022. Length control in abstractive summarization by pretraining information selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, Dublin, Ireland. Association for Computational Linguistics.

Timo Lohrenz, Björn Möller, Zhengyang Li, and Tim Fingscheidt. 2022. Relaxed attention for transformer models.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Takashi Maekaku, Yuya Fujita, Yifan Peng, and Shinji Watanabe. 2022. Attention Weight Smoothing Using Prior Distributions for Transformer-Based End-to-End ASR. In *Proc. Interspeech 2022*, pages 1071–1075.

Matthew Mccrudden, Joe Magliano, and G. Schraw. 2011. Relevance in text comprehension. *Text Relevance and Learning from Text*, pages 1–17.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

OpenAI. 2023. Gpt-4 technical report.

Mike Rosner and Carl Camilleri. 2008. MultiSum: Query-based multi-document summarization. In *Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization*, pages 25–32, Manchester, UK. Coling 2008 Organizing Committee.

Courtney Stevens and Daphne Bavelier. 2012. The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental Cognitive Neuroscience*, 2:S30–S48.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Follow the wisdom of the crowd: Effective text generation via minimum bayes risk decoding.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. PRIMER: pyramid-based masked sentence pre-training for multi-document summarization. *CoRR*, abs/2110.08499.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2022. Macsum: Controllable summarization with mixed attributes.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

A Details of the Datasets

NEWTS (Bahrainian et al., 2022) is a topic-oriented summarization dataset, for each article, there are two human-written summaries regarding two different topics, represented by a sequence of topic words and phrases, and a topic sentence. When building the dataset, they first obtain the topics using the LDA model (Blei et al., 2003) on the CNNDM dataset (Nallapati et al., 2016), and select the articles with two strong coherent topics, then the human annotators are asked to write two summaries regarding both topics (given by the top frequency words and manually written phrases) for each article. Based on how the dataset is built, the topic words and phrases do not necessarily appear in both the source article and human-written summaries. For simplicity, we only use the topic words in our experiments.

EntSUM (Maddela et al., 2022) is an entity-centric summarization dataset sampled from NYT dataset. Each data example contains a document, an entity extracted from the document, and a human-written summary regarding the given entity. The given entity is not necessarily central in the source document. As the original dataset is test-only, for the purpose of hyper-parameter search and few-shot experiments, we randomly split the original dataset into train/test sets.

We show two examples from the two datasets in Table 7

B Details of the Baseline Models

To evaluate the effectiveness of our method, we use three pre-trained models as the backbone model, including a pre-trained model for general summarization (PEGASUS), a model fine-tuned for the general summarization task (BART-CNNDM), and a model trained for conditional summarization task (CTRLsum). As both PEGASUS and BART-CNNDM are (pre-)trained for the general summarization task, we prepend the source documents with the given controlling aspects as prompts to fit the model to the controlled summarization task, followed by a special token and the source document as the inputs to the models. Besides, we also compare our method with the three backbone models, when they are extended with a widely used plug-and-play module (PPLM) (Dathathri et al., 2020).

PEGASUS (Zhang et al., 2020) is a pre-trained generative model tailored for abstractive summarization, with the objective of generating the mask-out salient sentences in the source documents. It shows a good performance for the general summarization task in zero-shot and few-shot settings.

BART-CNNDM (Lewis et al., 2020) is a pre-trained generative language model, fine-tuned on the CNNDM dataset (Nallapati et al., 2016) for the general summarization task, which achieves the state-of-the-art on the CNNDM dataset.

CTRLsum (He et al., 2020) is a pre-trained model for guided summarization, which is trained with the oracle keyphrases on the CNNDM dataset (Nallapati et al., 2016). Specifically, in the training stage, the oracle keyphrases (the matched keyphrases between the source document and the ground truth summary) is prepended to the source documents as the input, and in the inference stage, the keywords can either be automatically generated or human given.

PPLM (Dathathri et al., 2020) is a plug-and-play language model for conditional generation, which can be applied to any transformer-based generative model. It combines the pre-trained generative language model with one or more trained simple attribute models that guide text generation without any further training of the LM. Specifically, at each generation step, it updates the language model's hidden states using the attribute model's gradients with the current generated text as the input. The updates toward the direction to the combination of higher log-likelihood (LL) of the attribute a under the conditional attribute model $P(a|x)$ and higher LL of the unmodified language model $P(x)$. In this work, as we focus on content-controlling, we simply use the bag-of-words of the controlling factor as the attribute model. We apply the plug-and-play component to all the aforementioned models. As the PPLM module can not be trained, we only evaluate them in the zero-shot setting.

C Settings and Implementations

For the PPLM model, we do a hyper-parameter search on a randomly sampled validation set with size 100 for all the backbone models on both datasets. The hyper-parameters are γ_{gm} and step size, which are searched within $\{0.65, 0.75, 0.85, 0.95\}$ and $\{e^{-2}, e^{-3}, e^{-4}, e^{-5}\}$, respectvely.

NEWTS
Source: An American tourist has spent the night stranded in the Blue Mountains, west of Sydney, after she fell 15 metres off a cliff while bushwalking ... 'But due to the terrain in the prevailing weather that plan was aborted.' Rescue teams had to wait for the fog to lift so they could winch the woman out via a helicopter.
Summary: Foggy weather conditions made it difficult to rescue a stranded hiker. Helicopters cannot fly with such low visibility. The weather also blocked out sunlight.
Controlling Aspects: snow, weather, cold, winter, temperatures, conditions, hot, morning, expected, parts
EntSUM
Source: A lieutenant colonel in the Army Reserve, a firefighter and a college student are among 24 Manhattan and Long Island men accused of soliciting sex from detectives posing as children in computer chat rooms during a monthlong sting operation. ... The reporter, James Buckley, 50, of Brookhaven, who has worked for more than 20 years at WALK-FM (97.5), was arraigned on Tuesday morning in United States District Court in Central Islip. Magistrate Judge Arlene R. Lindsay ordered him held under house arrest on $350,000 bond. ... Commissioner Dormer said the detectives stayed in character, as naïve children. "The predators seduced the youngsters; the youngsters did not seduce the predators," he said.
Summary: The reporter, James Buckley, 50, of Brookhaven, who has worked for more than 20 years at WALK-FM (97.5), was arraigned on Tuesday morning in United States District Court in Central Islip. ... Mr. Buckley faces up to 19 years in federal prison if convicted.
Controlling Aspects: James Buckley

Table 7: Examples from both datasets.

For all the models, we use the default settings for the generation step as in the original model, and the length limit of the decoder is set to $56/142$, $86/172$ for NEWTS and EntSUM respectively based on the average length of the two datasets.

For the few-shot experiments, we train all the model on 10 data examples for 5 times (each time with different sampled examples). We use Adam(Kingma and Ba, 2015) as the optimizer, with the learning rate is 3e-5, and the models are trained for at most 100 epochs with early stops. We train the models on single A100 GPU, with the average time for training each model being 2-4 hours.

D Qualitative Analysis

In Table 8, we show the full example as shown in Table 5. It can be found that the PEGASUS model directly copies the first sentence (along with the prompts) to the summary, and with RelAttn, the model additionally generates a relevant sentence.

E Human Evaluation

Specifically, we randomly select 40 examples (20 from each dataset) with two system summaries generated by BART-CA+DOC, BART-RelAttn(OS) and the ground-truth summaries, and 3 annotations are collected for each example. For each data example, we show the annotators the original documents, the controlling aspects, and the three summaries (in a random order), and we ask the annotators to score each summary, regarding Relatedness, Factualness, Grammarticality, Coherence and the Overall quality within range 1-7, with 1 and 7 indicating very poor and very good for each aspect. The results are shown in Table 6. [11]

We show the annotation page for mechanical turkers in Figure 3,4,5.

F Software and Licenses

Our code is licensed under Apache License 2.0. Our framework dependencies are:

- HuggingFace Datasets[12], Apache 2.0
- NLTK [13], Apache 2.0
- Numpy[14], BSD 3-Clause "New" or "Revised"
- Transformers[15], Apache 2.0
- Pytorch[16], Misc
- Pytorch Lightning [17],Apache 2.0
- ROUGE [18], Apache 2.0
- BART [19], Apache 2.0

[11] We refer the questions for Grammaticality and Coherence from DUC human evaluation guidelines https://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt

[12] https://github.com/huggingface/datasets/blob/master/LICENSE

[13] https://github.com/nltk/nltk

[14] https://github.com/numpy/numpy/blob/main/LICENSE.txt

[15] https://github.com/huggingface/transformers/blob/master/LICENSE

[16] https://github.com/pytorch/pytorch/blob/master/LICENSE

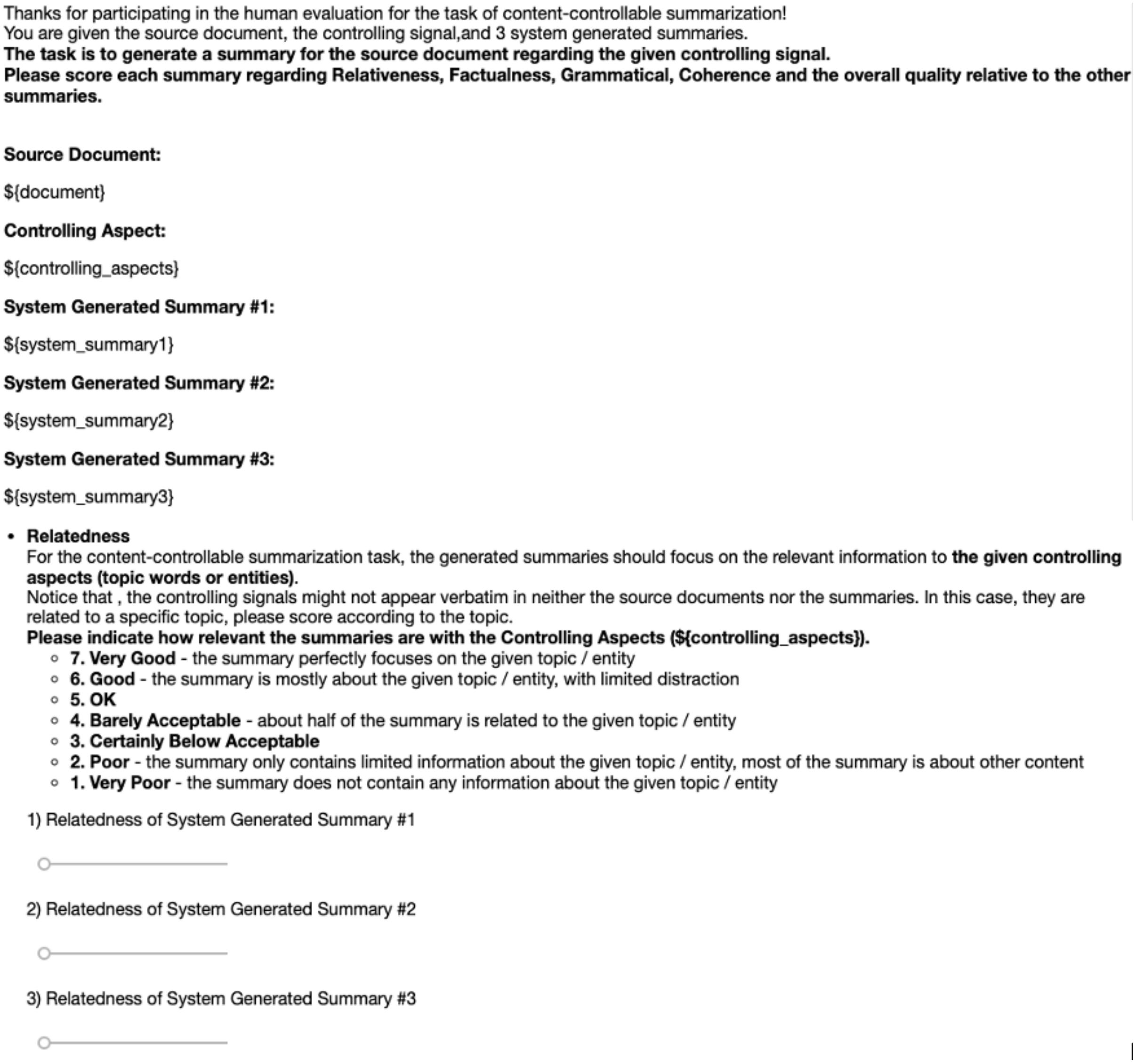
[17] https://github.com/PyTorchLightning/pytorch-lightning/blob/master/LICENSE

[18] https://github.com/google-research/google-research/tree/master/rouge

[19] https://huggingface.co/facebook/bart-large-cnn

Model Type	Summary
gt	The reporter, James Buckley, 50, of Brookhaven, who has worked for more than 20 years at WALK-FM (97.5), was arraigned on Tuesday morning in United States District Court in Central Islip. Allen Bode, an assistant United States attorney, said in a telephone interview that Mr. Buckley was caught trading pornography, including some hard-core material involving toddlers, with undercover agents in Germany and the United States. Mr. Buckley faces up to 19 years in federal prison if convicted.
PEGASUS-DOC	A lieutenant colonel in the Army Reserve, a firefighter and a college student are among 24 Manhattan and Long Island men accused of soliciting sex from detectives posing as children in computer chat rooms during a monthlong sting operation, the Suffolk County police said on Tuesday. ”Let this be a wake-up call to parents to monitor our children’s chat rooms,” said County Executive Steve Levy at a news conference at police headquarters in Yaphank. The Suffolk police commissioner, Richard Dormer, who was also at the news conference, said, ”I consider this the sleaziest kind of investigation, but I consider it necessary.” In a separate case, a traffic and weather reporter for a Long Island radio station was indicted on federal charges of possessing child pornography. Detectives in the Police Department’s computer crimes unit who were at the
PEGASUS-CA+DOC	James Buckley => A lieutenant colonel in the Army Reserve, a firefighter and a college student are among 24 Manhattan and Long Island men ... (same as above)
PEGASUS+RelAttn(OS) $(w_{rel} = 0.10)$	A lieutenant colonel in the Army Reserve, a ...(same as above) federal charges of possessing child pornography. Buckley Buckley Buckley was caught trading pornographic messages, including some hard-core
BART-Doc	24 men accused of soliciting sex from detectives posing as children in computer chat rooms. In separate case, a traffic and weather reporter for a Long Island radio station indicted on federal charges of possessing child pornography. All were charged with attempting to disseminate indecent material to a minor, a felony with a maximum penalty of four years in prison. The 24 cases together constitute the largest child sex crackdown ever in Suffolk, the police said.
BART-CA+DOC	24 men accused of soliciting sex from detectives posing as children in computer chat rooms. Police commissioner: ”I consider this the sleaziest kind of investigation, but I consider it necessary” In separate case, Long Island radio reporter indicted on federal charges of possessing child pornography.
BART+RelAttn (OS-Best) $(w_{rel} = 0.12)$	Buckley faces up to 19 years in federal prison if convicted.Buckley has worked for more than 20 years at WALK-FM (97.5) Buckley’s court-appointed defender, Randi Chavis, said in a telephone interview that he pleaded not guilty. Buckley was caught trading pornography, including some hard-core material involving toddlers, with undercover agents in Germany and the United States.Buckley’s attorney declined to comment.
BART+RelAttn (OS) $(w_{rel} = 0.15)$	Buckley faces up to ...*(same as above)* declined to comment.Buckley is one of 11 men arrested after sending sexually explicit messages.
BART+RelAttn (Val) $(w_{rel} = 0.17)$	Buckley faces up to ...*(same as above)* undercover agents in Germany and the United States.Buckley’s attorney, Buckley Buckley, declined to comment.Buckley is one of 11 men arrested after sending sexually explicit messages.
BART+RelAttn $(w_{rel} = 0.30)$	Buckley Buckley, Buckley Buckley Buckley is Buckley Buckley’s Buckley Buckley.Buckley Buckley faces Buckley Buckley was ...
CTRLsum	James Buckley, 50, of Brookhaven, faces up to 19 years in federal prison if convicted. Police say they created online profiles and screen names for fictitious boys and girls. Police commissioner: ”I consider this the sleaziest kind of investigation, but I consider it necessary” Police say 11 others were arrested after sending sexually explicit messages or photographs online; their identities were provided by the Internet service providers under subpoena. autoimmune autoimmune autoimmune disease. autoimmune disease, autoimmune disease of the mind, immune system, and body.
CTRLsum+RelAttn(OS) $(w_{rel} = 0.05)$	James Buckley, 50, of Brookhaven, faces up to 19 years in federal prison if convicted. Buckley was caught trading pornography, including hard-core material involving toddlers, with undercover agents in Germany and the U.S. Police say they created online profiles and screen names for fictitious boys and girls, ages 12 to 14, and visited chat rooms operated by AOL, Yahoo and other Internet service providers. Police commissioner: ”I consider this the sleaziest kind of investigation, but I consider it necessary”

Table 8: The full example from the EntSUM dataset (*doc_id: 1739833*) with the entity **James Buckley**.



Thanks for participating in the human evaluation for the task of content-controllable summarization!
You are given the source document, the controlling signal,and 3 system generated summaries.
The task is to generate a summary for the source document regarding the given controlling signal.
Please score each summary regarding Relativeness, Factualness, Grammatical, Coherence and the overall quality relative to the other summaries.

Source Document:

${document}

Controlling Aspect:

${controlling_aspects}

System Generated Summary #1:

${system_summary1}

System Generated Summary #2:

${system_summary2}

System Generated Summary #3:

${system_summary3}

- **Relatedness**
  For the content-controllable summarization task, the generated summaries should focus on the relevant information to **the given controlling aspects (topic words or entities).**
  Notice that , the controlling signals might not appear verbatim in neither the source documents nor the summaries. In this case, they are related to a specific topic, please score according to the topic.
  Please indicate how relevant the summaries are with the Controlling Aspects (${controlling_aspects}).
  - **7. Very Good** - the summary perfectly focuses on the given topic / entity
  - **6. Good** - the summary is mostly about the given topic / entity, with limited distraction
  - **5. OK**
  - **4. Barely Acceptable** - about half of the summary is related to the given topic / entity
  - **3. Certainly Below Acceptable**
  - **2. Poor** - the summary only contains limited information about the given topic / entity, most of the summary is about other content
  - **1. Very Poor** - the summary does not contain any information about the given topic / entity

  1) Relatedness of System Generated Summary #1

  ○

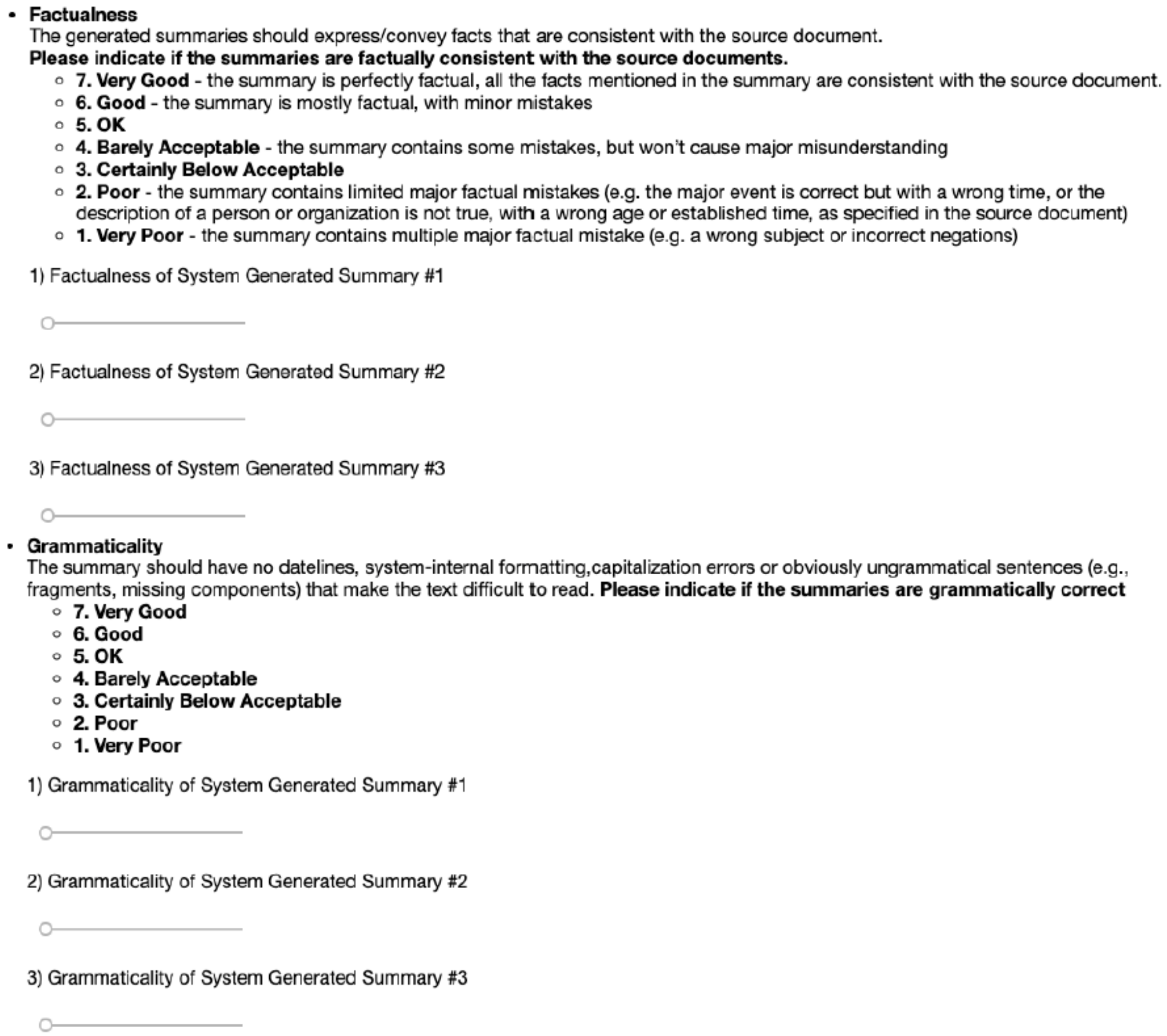
  2) Relatedness of System Generated Summary #2

  ○

  3) Relatedness of System Generated Summary #3

  ○

Figure 3: The questions for human evaluation.



- **Factualness**
  The generated summaries should express/convey facts that are consistent with the source document.
  Please indicate if the summaries are factually consistent with the source documents.
  - **7. Very Good** - the summary is perfectly factual, all the facts mentioned in the summary are consistent with the source document.
  - **6. Good** - the summary is mostly factual, with minor mistakes
  - **5. OK**
  - **4. Barely Acceptable** - the summary contains some mistakes, but won't cause major misunderstanding
  - **3. Certainly Below Acceptable**
  - **2. Poor** - the summary contains limited major factual mistakes (e.g. the major event is correct but with a wrong time, or the description of a person or organization is not true, with a wrong age or established time, as specified in the source document)
  - **1. Very Poor** - the summary contains multiple major factual mistake (e.g. a wrong subject or incorrect negations)

  1) Factualness of System Generated Summary #1

  ○

  2) Factualness of System Generated Summary #2

  ○

  3) Factualness of System Generated Summary #3

  ○

- **Grammaticality**
  The summary should have no datelines, system-internal formatting,capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read. **Please indicate if the summaries are grammatically correct**
  - **7. Very Good**
  - **6. Good**
  - **5. OK**
  - **4. Barely Acceptable**
  - **3. Certainly Below Acceptable**
  - **2. Poor**
  - **1. Very Poor**

  1) Grammaticality of System Generated Summary #1

  ○

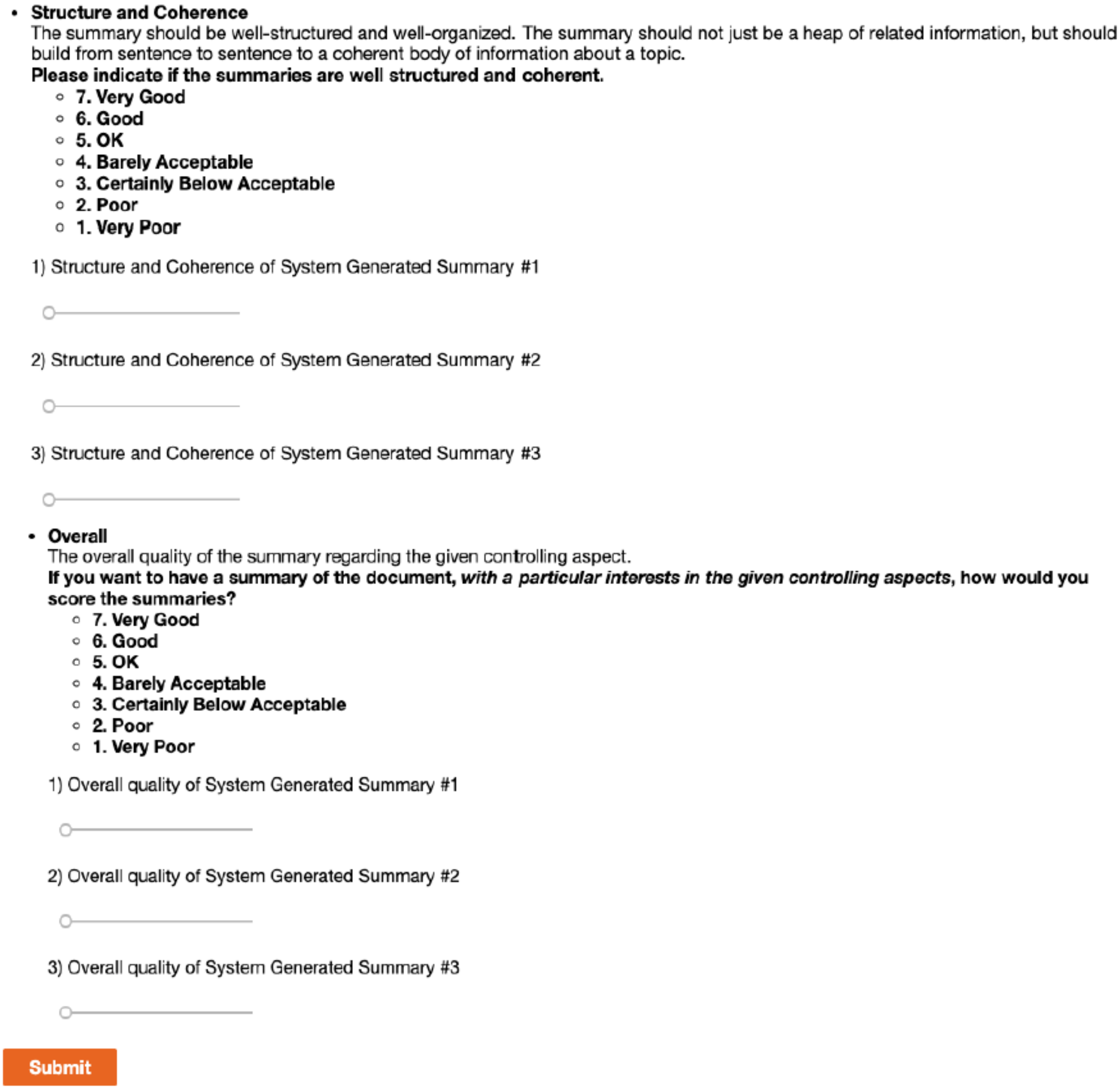
  2) Grammaticality of System Generated Summary #2

  ○

  3) Grammaticality of System Generated Summary #3

  ○

Figure 4: The questions for human evaluation.



- **Structure and Coherence**
  The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.
  Please indicate if the summaries are well structured and coherent.
  - **7. Very Good**
  - **6. Good**
  - **5. OK**
  - **4. Barely Acceptable**
  - **3. Certainly Below Acceptable**
  - **2. Poor**
  - **1. Very Poor**

  1) Structure and Coherence of System Generated Summary #1

  2) Structure and Coherence of System Generated Summary #2

  3) Structure and Coherence of System Generated Summary #3

- **Overall**
  The overall quality of the summary regarding the given controlling aspect.
  If you want to have a summary of the document, *with a particular interests in the given controlling aspects*, how would you score the summaries?
  - **7. Very Good**
  - **6. Good**
  - **5. OK**
  - **4. Barely Acceptable**
  - **3. Certainly Below Acceptable**
  - **2. Poor**
  - **1. Very Poor**

  1) Overall quality of System Generated Summary #1

  2) Overall quality of System Generated Summary #2

  3) Overall quality of System Generated Summary #3

Figure 5: The questions for human evaluation.

- PEGASUS[20], Apache 2.0
- CTRLSum[21], BSD-3-Clause license

[20] https://huggingface.co/google/pegasus-large

[21] https://huggingface.co/hyunwoongko/ctrlsum-cnndm