MACSUM: Controllable Summarization with Mixed Attributes

Yusen Zhang[1*] **Yang Liu**[2†] **Ziyi Yang**[2]
Yuwei Fang[2] **Yulong Chen**[3] **Dragomir Radev**[4]
Chenguang Zhu[2] **Michael Zeng**[2] **Rui Zhang**[1]
[1]Penn State University, USA [2]Microsoft Research, USA
[3]Westlake University, China [4]Yale University, USA
{yfz5488,rmz5227}@psu.edu; yaliu10@microsoft.com

Abstract

Controllable summarization allows users to generate customized summaries with specified attributes. However, due to the lack of designated annotations of controlled summaries, existing work has to craft pseudo datasets by adapting generic summarization benchmarks. Furthermore, most research focuses on controlling single attributes individually (e.g., a short summary **or** a highly abstractive summary) rather than controlling a mix of attributes together (e.g., a short **and** highly abstractive summary). In this paper, we propose MACSUM, the first human-annotated summarization dataset for controlling mixed attributes. It contains source texts from two domains, news articles and dialogues, with human-annotated summaries controlled by five designed attributes (Length, Extractiveness, Specificity, Topic, and Speaker). We propose two simple and effective parameter-efficient approaches for the new task of mixed controllable summarization based on hard prompt tuning and soft prefix tuning. Results and analysis demonstrate that hard prompt models yield the best performance on most metrics and human evaluations. However, mixed-attribute control is still challenging for summarization tasks. Our dataset and code are available at https://github.com/psunlpgroup/MACSum.

1 Introduction

Text summarization is the task of compressing the input text into a concise and coherent version by preserving salient information. There has been substantial progress in generic summarization by generating one overall summary for each input (McKeown and Radev, 1995; Erkan and Radev, 2004; Rush et al., 2015; Cheng and Lapata, 2016; See et al., 2017; Paulus et al., 2018). However, readers have diverse preferences when summarizing the same article, such as topics, speakers, or lengths of the summary (Fan et al., 2018; Zhong et al., 2021; Goyal et al., 2022b). Therefore, generating customized summaries to meet different preferences is a natural capability of summarization systems.

Due to the lack of a human-annotated controllable summarization benchmark, existing research has to adapt generic datasets to create pseudo-controllable summaries (Fan et al., 2018; He et al., 2020; Zhong et al., 2021; Goyal et al., 2022b; Chan et al., 2021). He et al. (2020), for example, extract topics from a generic summary by assuming the summary is controlled by the extracted topics to evaluate summarization over topics. However, this adaptation-based setting raises three issues. First, the adapted summaries are not really written with the guidance of being controlled by the designed attributes. Second, this conversion can only build one target summary for each source, while it is preferable to have summaries with different control attributes for the *same* input article. Third, for attributes like Extractiveness or Specificity, there are no straightforward adaptation methods.

Meanwhile, previous studies mostly focus on controlling a single attribute, e.g., generating a short summary **or** a highly abstractive summary. However, mixing different control attributes is more challenging and underexplored (Russo et al., 2020). For example, Figure 1 shows a case of mixed-attribute control by requiring a short summary regarding ‘‘Pope Francis’’, or a highly extractive and highly specific summary on the topic ‘‘blood moon’’. Users can simultaneously control different attributes in the generated summary.

[*] Yusen Zhang completed this work during his internship at Microsoft.

[†] Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 787–803, 2023. https://doi.org/10.1162/tacl_a_00575
Action Editor: Emily Pitler. Submission batch: 12/2022; Revision batch: 2/2023; Published 7/2023.

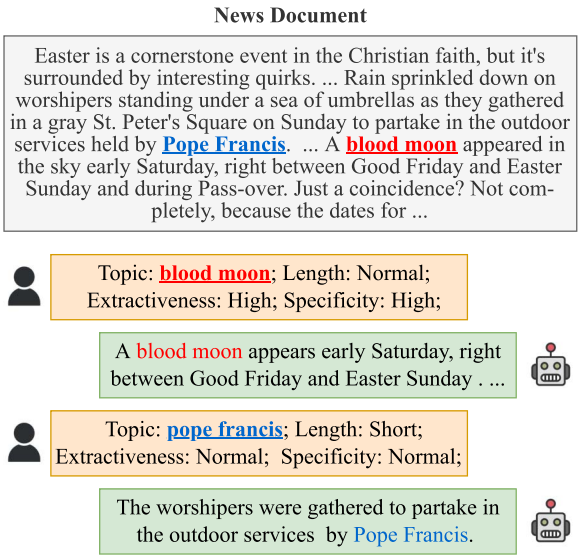


Figure 1: An example of MACSUM. For the same input source text, the system needs to generate different reference summaries (green boxes) for different mixed control attributes (orange boxes).

In this paper, we propose MACSUM, a human-annotated benchmark for controllable summarization with mixed attributes. In MACSUM, source texts are collected from both news and dialogue domains. We define five control attributes of summarization by synthesizing previous studies (Chan et al., 2021; Liu et al., 2018; Fan et al., 2018), including Length (*Len*), Extractiveness (*Ext*), summarization Specificity (*Spe*), Topic (*Tpc*), and Speaker (*Spk*).[1] For each input source, we sample a set of different combinations of these attributes for human annotations. The resulting MACSUM dataset contains a rich set of annotations of human-written summaries for the same input with different mixtures of control attributes. Table 1 compares MACSUM with previous work, and MACSUM is the first to mix these five attributes with human annotations, covering both dialogue and document source texts.

Furthermore, to establish a baseline of mixed-attribute control, we design two simple yet effective frameworks that can steer a summarization model by either hard prompt (HP) or soft prefix (SP) inspired by prompt learning (Raffel et al., 2020; Li and Liang, 2021). For each value of a control attribute (e.g., long length), in the HP framework, we prepend the description of control attributes (e.g., ‘‘Length: Long’’) to the input source as hard prompts; in the SP framework, we assign a set of external parameters, called soft prefixes, to the model. The summarization model is trained to summarize an input with hard prompt/soft prefixes of different control signals. We evaluate these baseline models on MACSUM with proposed two automatic evaluation metrics measuring the quality of control. Our results and analysis in two domains reveal that the HP framework yields the best performance on all automatic metrics and human evaluations, although mixed-attribute control is still challenging.

[1]The speaker attribute is for the dialogue domain only.

2 Related Work

2.1 Controllable Summarization

Previous work on controllable text summarization focuses on length (Fan et al., 2018; Liu et al., 2018; Makino et al., 2019; Saito et al., 2020; Liu et al., 2022; He et al., 2022; Goyal et al., 2022a), entity (He et al., 2020; Narayan et al., 2021; Maddela et al., 2022; Hofmann-Coyle et al., 2022), aspect (Tan et al., 2020; Amplayo et al., 2021), content (Dou et al., 2021; Xiao et al., 2022), style (Cao and Wang 2021), granularity (Zhong et al., 2022), and abstractiveness (Song et al., 2020). Query-focused summarization (Dang, 2005; Fisher and Roark, 2006; Daumé III and Marcu, 2006) generates summaries for specific user information requests, but it does not explicitly control the output style. Furthermore, interactive summarization (Bornstein et al., 1999; Leuski et al., 2003) and reinforcement learning guided summarization (Paulus et al., 2018; Böhm et al., 2019; Stiennon et al., 2020) have been used to incorporate human preferences and feedback, yet the human feedback explored so far is largely limited to the generic quality of summaries instead of fine-grained attributes. Notably, Chan et al. (2021) propose a constrained Markov Decision Process for controllable summarization for different attributes, but it is unclear if it can perform multi-attribute control. Goyal et al. (2022b) investigate multi-feature control by mixing multiple decoders, yet their solution is only based on decoding improvements that yield suboptimal controlling performance. Therefore, most previous work is over-specialized for controlling particular attributes, while controlling multiple attributes is still underexplored. Further more, existing works are mostly evaluated on

	Domain	Source Type		Construction		Mixed Attr.	Control Attributes				
		Dial.	Doc.	Anno.	Multi-O.		*Tpc*	*Spk*	*Len*	*Ext*	*Spe*
CASum (Fan et al., 2018)	News	✗	✓	✗	✗	✗	✓	✓	✓	✗	✗
CTRLSum (He et al., 2020)	News, Papers	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗
QMSum (Zhong et al., 2021)	Meetings	✓	✗	✓	✓	✓	✓	✓	✗	✗	✗
HydraSum (Goyal et al., 2022b)	News	✗	✓	✗	✗	✓	✗	✗	✓	✓	✓
CMDP (Chan et al., 2021)	News	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗
MACSUM (ours)	News, Meetings	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between MACSUM and previous work on controllable summarization. Dial. and Doc. means if the source is dialogue or document. Anno. indicates whether the data is constructed by human annotation or rule-based pseudo-split. Multi-O. shows if there are multiple outputs with different control attributes for the same source. Mixed Attr. shows if mixed attribute control is allowed. Control Attributes are defined in Section 3.

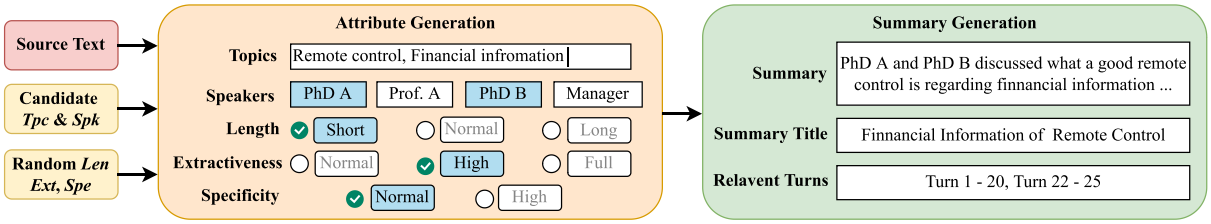


Figure 2: Annotation pipeline of MACSUM. The annotator needs to summarize the contents of meetings/documents according to the five control attributes, give the relevant text spans, and write a summary title.

pseudo datasets adapted from generic summarization datasets.

2.2 Prompt Learning

Prompt learning is first proposed in GPT-3 (Brown et al., 2020), where large pretrained language models can perform desired tasks with the guidance of instructions and examples. Some efforts explore prompt-tuning using natural language by converting original inputs into cloze-style questions and then tuning language models (Shin et al., 2020; Schick and Schütze, 2021; Chen et al., 2022; Min et al., 2022). However, most of them focus on natural language understanding tasks and usually need a careful selection of prompts. Instead of using human-crafted tokens, other work explores using continuous vectors as prompts (Lester et al., 2021; Qin and Eisner 2021; Liu et al., 2021; Li and Liang, 2021). Among them, prefix-tuning is particularly designed for text generation (Li and Liang, 2021). Prefix-tuning prepends trainable vectors to each layer of language models as prefixes and keeps other parameters frozen during training. In this work, we propose two methods for mixed attribute controllable summarization based on prompt-tuning and prefix-tuning, respectively.

3 The MACSUM Dataset

To provide a benchmark for controllable summarization, we propose MACSUM, a high-quality human-annotated mixed-attribute controlled summarization dataset. Inspired by several previous studies on controllable generation (Chan et al., 2021; Liu et al., 2018; Fan et al., 2018), MACSUM is annotated with 5 types of control attributes, including Topic, Speaker, Length, Extractiveness, and Specificity (Section 3.1).

As shown in Figure 1, these five attributes can be combined together in various designs (Section 3.1). Additionally, Topic and Speaker can have multiple values as well, i.e., more than one speaker or topic to focus on. In annotation, we require the corresponding summary to fulfill all attributes together.

The data annotation pipeline is divided into four steps (Figure 2). First, we carefully select the source texts from several widely used summarization datasets in news or dialogue domains. Second, some automatic tools are leveraged to form a pool of candidate attributes as guidance for the next step. Third, the annotators manually label five attributes to form a control attribute set and repeat the process multiple times. Finally, the

annotators write down the summary that meets each control attribute set.

3.1 Control Attributes

MACSUM provides five attributes to control the summary generation.

Topic (*Tpc*) indicates certain contents of the source that users are particularly interested in. The summary should only contain contents that are related to the given topics. We provide multiple keywords such as ''remote control, financial information'' for annotators as candidate topics.

Speaker (*Spk*) indicates certain speakers in a dialogue whose content is preferred by the user. In MACSUM (MAC-Dial only), this is specified by giving the name of certain speakers, such as ''Program Manager''.

Length (*Len*) indicates the number of words of the summary, serving the time budget for users to read this summary. In MACSUM, *Len* is controlled by [*short*, *normal*, *long*], three values of this attribute.[2] Our annotation guideline provides a reference range of compression ratio and word count for each length value.

Extractiveness (*Ext*) describes the proportion of the summary extracted from the source text. This is useful when users sometimes want content directly extracted from the source, while sometimes they want more abstractive and more readable results. In MACSUM, *Ext* can take values of [*normal*, *high*, *full*].

Specificity (*Spe*) means how many details or descriptive contents we need to include in the summary. Referring to Louis and Nenkova (2011), different users can prefer more general summaries or more specific summaries. MACSUM contains two levels of *Spe* control, namely [*normal*, *high*], where normal is the natural specificity and high requires more specific contents.

Specificity differs from Length. Length is the number of words, while Specificity is the density of descriptive contents (e.g., numbers, entities, and names). Thus, a short summary can have a higher Specificity than a long one.

MACSUM supports Mixed-Attribute Control because it is a natural need for users to control multiple aspects at the same time, e.g., wanting the summary to be short, highly extractive, and only talking about some topics. To this end, as shown in Figure 1, the samples in MACSUM can control multiple attributes simultaneously. We require the annotated summaries to meet all requirements at the same time. If some combinations are considered too difficult to fulfill, we allow annotators to skip them in rare cases. We provide detailed distribution of attributes in Figure 3.

[2]We denote attribute value as *Attribute: value*—e.g., *Len: long*.

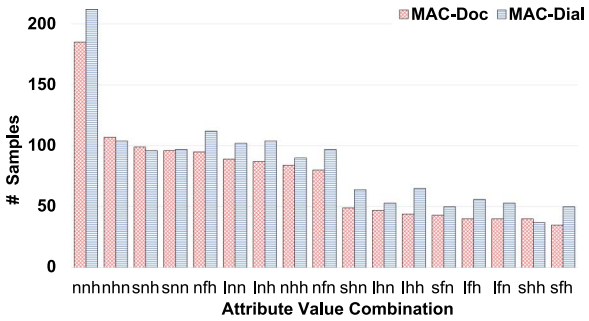


Figure 3: Distribution of mixed attributes. Each category is represented by the first character of its control attribute values, e.g., *snh* represents *Len: short*, *Ext: normal*, and *Spe: high*.

3.2 Annotation Pipeline

Source Selection MACSUM covers source text from both document and dialogue summarization tasks. We pick **CNNDM** (Hermann et al., 2015) as the document dataset and **QMSum** (Zhong et al., 2021) as the dialogue dataset. CNNDM is a large-scale document summarization dataset containing news stories along with their corresponding highlights, collected from CNN and Daily Mail websites. QMSum is a popular query-based meeting summarization dataset. It contains the transcripts of three domains, including AMI, ICSI, and committee meetings of the Welsh Parliament and the Parliament of Canada. For CNNDM, we randomly pick 10k documents in the test set for the annotation. For QMSum, we first split each meeting into shorter units according to the topic partition and discard the units that are longer than 5000 tokens.

Attribute Candidate Extraction For Topic, we first use a keyword extraction tool (Boudin, 2016) to extract the top 20 keywords from the source text as candidates. For Speaker, we collect all speakers in the source text to form a candidate

set. For the remaining Length, Extractiveness, and Specificity attributes, we generate their values and combination randomly from a uniform distribution, mimicking the behavior of users with diverse needs for customized summaries.

Attribute Generation We hire 4 native English speakers as annotators. The annotators can either freely choose topics from the candidate topics or write the keywords by themselves. As for the Speaker attribute, we ask the annotators to pick one or more names from the candidate set. Besides, Length, Extractiveness, and Specificity are automatically filled with randomly generated values.

Attribute generation repeats several times for each source to form various attribute combinations, so-called samples. Overall, each source text contains eight samples for every 2000 words.

Summary Generation We first ask all annotators to read our annotation guideline and 10 annotated examples. Afterward, given several combinations of control requirements, i.e., the control attribute sets, the annotators follow our guidance and write a summary for each control combination.

We also ask them to annotate the related text spans for use in future work, such as retrieval-based methods. Related text spans are the turns/sentences in the source that are most relevant to the golden summary. These spans are the minimum necessary turns/sentences the annotators need to produce the complete summary.

Finally, the annotators read the summary again for quality assurance, and we ask them to write a short title for this summary, e.g., ''discussion of remote control style''. This is helpful for future work such as title generation, and it also provides us with a quick way to verify whether the annotators read their generated summaries.

Quality Control First, we control the annotation quality through a careful pilot test. Before the annotation process starts, annotators are carefully selected via a pilot test. We assign each annotator the same three input texts with various mixed attributes, and we choose the qualified annotators according to annotation results.

Second, we conduct a weekly sampling inspection. We frequently monitor the quality of annotations. We collect the results weekly and provide feedback to the annotators to ensure quality.

3.3 Automatic Metrics

Overview Along with the annotated benchmark, we also design a system of automatic metrics for evaluating the model's capability to generate controllable summaries. For each attribute, we define its own attribute metric function to represent the degree of control. We then propose **Control Error Rate** (CER) and **Control Correlation** (CC). CER measures the distance between the generated and golden summary in terms of their degrees of control using attribute metric functions. A good model should have smaller CER $\downarrow$. CC measures the distribution of attribute metric functions among generated summaries with different attribute values, representing the model's capability to correlate to the definition of the control attribute. A good model should have a CC distribution that is similar to that of the golden summary $\updownarrow$. In addition, we also report F-1 of ROUGE-1/2/L (Lin, 2004) for the general quality of the summary $\uparrow$.

Definition For a control attribute r and its attribute metric function f_r, given a predicted summary $\hat{y}$, golden summary y, Control Error Rate (CER) is defined as:

$$\text{CER}(\hat{y}, y) = \frac{|f_r(\hat{y}) - f_r(y)|}{f_r(y) + \epsilon} \tag{1}$$

where ϵ is a small value to avoid error when $f_r(y)$ is zero.

Additionally, for the control attribute r (e.g., *Len*) with a control value pair $[v_1, v_2]$ (e.g., [*short*, *long*]), predicted summaries for these two values $[\hat{y}_1, \hat{y}_2]$, Control Correlation (CC) is defined as:

$$\text{CC}(\hat{y_1}, \hat{y_2}) = \frac{f_r(\hat{y}_1) - f_r(\hat{y}_2)}{\text{Distance}(v_1, v_2)} \tag{2}$$

where $\text{Distance}(v_1, v_2)$ calculates the distance *from* control value v_1 *to* v_2, which can be negative. For instance, $\text{Distance}(\textit{high}, \textit{normal}) = 1$, and $\text{Distance}(\textit{short}, \textit{long}) = -2$. When CC is above/below 0, it indicates the evaluated model has a positive/negative correlation with the control objective. Additionally, CER and CC for multiple samples are their arithmetic mean.

For each of the five control attributes, we define its own attribute metric f_r which maps the summary to a real number that represents the degree of control. **Topic** f_{Tpc} is the proportion of topic

	#Samples/#Sources			Avg. Number in Text			Avg. # C.A.	
	Train	Dev	Test	Source Len.	Source Turns	Reference Len.	Topic	Speaker
CNNDM	2887k/2887k	13k/13k	11k/11k	781.0	–	56.0	–	–
QMSum	1257/162	272/35	279/35	9069.8	556.8	69.6	–	–
MAC-Doc	4278/755	554/94	547/94	835.4	–	54.1	0.8	–
MAC-Dial	2338/328	292/41	324/41	2754.3	144.6	69.4	1.7	1.2

Table 2: Statistics of MACSUM consisting of two parts: MAC-Doc from CNNDM and MAC-Dial from QMSum. Source Len., Ref. Len. are tokens in source and reference. Topic, Speaker are the averaged number of topics/speakers.

keywords shown in the summary. **Speaker** f_{Spk} is the number of tokens spoken by the selected speakers divided by the total number of tokens in the summary. **Length** f_{Len} is the number of tokens in the summary. **Extractiveness** f_{Ext} is the average of ROUGE-2 precision and ROUGE-3 precision (Lin, 2004) of the generated summary against the input. For **Specificity**, inspired by previous studies (Resnik, 1995; Amplayo et al., 2021), we find that verb, noun, numeral, and the total number of tokens show the most significant information about specificity. Thus, we define $f_{Spe} = (0.1 \times vb + 0.2 \times tok + 0.3 \times nn + 0.4 \times cd)/n_s$, where vb, tok, nn, cd, and n_s represent the number of verbs, tokens, nouns, numeral tokens, and the number of sentences in the summary.

3.4 Statistics of MACSUM

Dataset Split and Source Data Distribution Table 2 shows the statistics. MACSUM covers two domains (MAC-Doc for news and MAC-Dial for dialogue) with 8333 annotated summaries (5379 in MAC-Doc and 2954 in MAC-Dial), paired with 1353 source inputs (943 in MAC-Doc and 410 in MAC-Dial). The averaged number of tokens in sources of MAC-Doc is shorter than that in the original QMSum dataset since we truncate the input into segments. We split the source text randomly into training/valid/test sets with 80%/10%/10%.

Distribution of Control Attribute Metrics With definitions from Section 3.3, Table 3 calculates automatic attribute metrics for all 5 control attributes. As presented, the annotated summaries with different control attribute values can distinguish from each other by a large margin. For example, samples with *Len: long* have a much longer input, and samples with *Ext: full* have a higher extractiveness metric. This verifies the high annotation quality of MACSUM and also proves that our proposed attribute metrics are consistent with the control objective of each control attribute.

Attribute	Value	MAC-Dial			MAC-Doc		
		Train	Dev	Test	Train	Dev	Test
Length	short	38.04	39.52	43.84	31.97	33.37	34.30
	normal	67.47	72.34	69.68	46.63	45.15	47.92
	long	104.03	93.37	107.44	92.37	90.74	95.35
Extractiveness	normal	0.26	0.28	0.23	0.29	0.29	0.27
	high	0.32	0.33	0.31	0.39	0.39	0.46
	full	0.49	0.43	0.50	0.63	0.63	0.61
Specificity	normal	5.25	4.90	5.01	4.73	4.70	4.67
	high	6.32	6.28	6.17	4.88	5.11	4.82
Topic	–	0.83	0.81	0.79	0.95	0.98	0.95
Speaker	–	0.74	0.71	0.71	–	–	–

Table 3: Attribute metric functions f_r of different control attribute values.

Mixed-Attribute Distributions Figure 3 shows the ratio of different combinations of the control attributes. This illustrates diverse combinations of mixed-attributes summaries by controlling *Len*, *Ext*, and *Spe* together in one sample.

4 Methods

For setting baseline results on MACSUM, we propose three models following previous research on controllable text generation using prompt learning. With the same input and different prompts, the large pretrained model is able to generate different results for different tasks, such as summarization and translation (He et al., 2020; Fan et al., 2018; Raffel et al., 2020). As shown in Figure 4, we leverage two types of prompt learning approaches to control the attributes of summaries, namely, hard prompt (HP) and soft prefix tuning (SP). We also test the combination of them, HP+SP.

Hard Prompt (HP) uses the description of control attributes as the hard prompt. Each attribute is

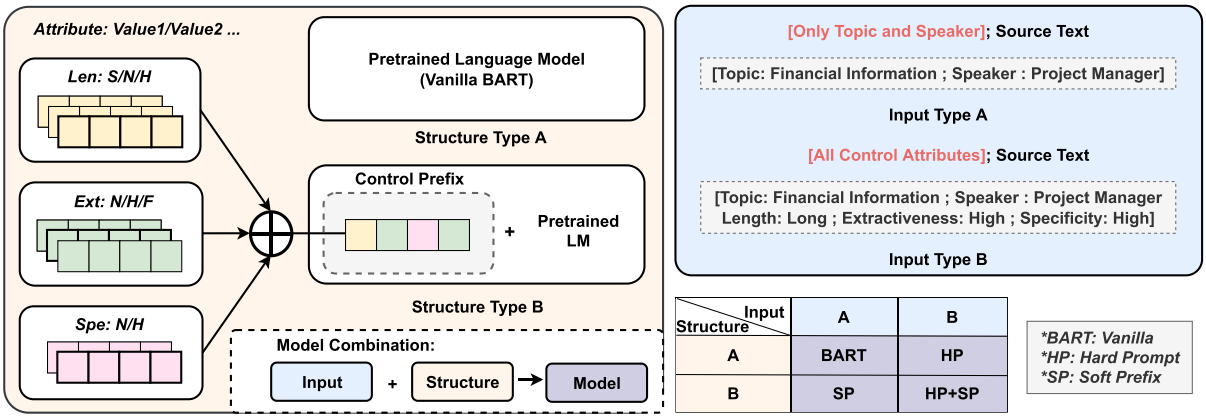


Figure 4: Comparison of different frameworks. For the HP model, the control attributes are prepended to the input to form a hard prompt. For the SP model, the selected prefix vectors are added together to form a control prefix. HP+SP contains both hard prompts and control prefixes.

formed as ''Attribute: Value'', where ''Attribute'' can be ''Topic, Speaker, Length, Extractiveness, Specificity'', and ''Value'' is the corresponding value (e.g., High or Normal) of the attribute. We concatenate 5 control attributes using '';'' and prepend it to the input source.

Soft Prefix (SP) follows Li and Liang (2021). We prepend external trainable parameters to both the encoder and decoder to control the summarization model. For controlling *Len*, *Ext*, and *Spe*, we assign m prefix embeddings for each attribute value where m is a hyper-parameter meaning the length of prefix, i.e., prefix length. Readers can refer to Li and Liang (2021) for implementation details. For example, for *Len: Long*, we assign $\boldsymbol{E}_{Len:long} = [\boldsymbol{e}^1_{Len:long}, \cdots, \boldsymbol{e}^m_{Len:long}]$ where $\boldsymbol{e}^j_i$ is a vector with dimension of word embedding. And for controlling an input case with a set $\mathcal{V}$ of mixed requirements, we sum the embeddings of all control attribute values: $\boldsymbol{E} = [\sum_{v\in\mathcal{V}} \boldsymbol{e}^1_v, \cdots, \sum_{v\in\mathcal{V}} \boldsymbol{e}^m_v]$. And for controlling *Tpc* and *Spk*, we use the embeddings of input topics words $\boldsymbol{E}_{Tpc}$ and input speaker names $\boldsymbol{E}_{Spk}$. This list of embedding vectors $\boldsymbol{E}$ is then prepended to each layer of the Transformer-based summarization model as external key/value vectors in its self-attention operations. $\boldsymbol{E}_{Tpc}$ and $\boldsymbol{E}_{Spk}$ are prepended only to the input layer.

Hard Prompt + Soft Prefix (HP+SP) combines both approaches by prepending the hard prompt of five attributes in HP and using prefix tuning in SP.

5 Experiments

In this section, we present the implementation details, experimental results, and human evaluation of models on MACSUM dataset.

5.1 Implementation Details

We use PyTorch and the HuggingFace library (Wolf et al., 2019) to implement our model. The experiments are conducted on 8 A100 GPUs.

We use BART (Lewis et al., 2020) as the backbone model. We also use a vanilla BART trained without control attribute input as a weak baseline (Appendix A). If not mentioned, we initialize the backbone using BART-large-cnn and then fine-tune it on the MACSUM dataset. We pick the $3e-5$ learning rate searching from $\{1e-5, 3e-5, 1e-4\}$. Additionally, n-gram blocking is set to 3, and we use the AdamW optimizer with 500 warmup steps. Dialogue inputs are flattened by separating turns with ''<\s>'' which we find yields better results.

5.2 Experiment Results

As mentioned in Section 3.3, we calculate Control Error Rate (CER) and Control Correlation (CC) metrics for evaluating control quality, and we also report ROUGE scores for evaluating summarization quality. For a model, its performance is better when the CER value is lower↓, ROUGE is higher↑, and its CC is closer to the golden summary↕.

Table 4 shows the results of MAC-Doc. The HP model obtains the highest performance on both

	Length		Extractiveness		Specificity		Topic	Average	Quality		
	CER↓	CC↕	CER↓	CC↕	CER↓	CC↕	CER↓	CER↓	R1↑	R2↑	RL↑
Gold	0.000	32.444	0.000	0.141	0.000	0.103	0.000	0.000	1.000	1.000	1.000
BART	0.486	0.000	1.177	0.000	0.490	0.000	0.345	0.624	0.290	0.102	0.250
HP	**0.340**	**31.421**	**0.802**	0.239	**0.353**	0.259	**0.333**	**0.457**	**0.300**	**0.104**	**0.261**
SP	0.475	4.671	1.111	0.055	0.466	**0.105**	0.471	0.631	0.261	0.092	0.228
HP+SP	0.373	25.226	1.136	**0.133**	0.370	0.191	0.358	0.559	0.288	0.103	0.248

Table 4: Results on MAC-Doc. The performance of the model is better when Control Error Rate (CER) is lower ↓, ROUGE is higher ↑, and Control Correlation (CC) is closer to the golden summary ↕.

	Length		Extractiveness		Specificity		Topic	Speaker	Average	Quality		
	CER↓	CC↕	CER↓	CC↕	CER↓	CC↕	CER↓	CER↓	CER↓	R1↑	R2↑	RL↑
Gold	0.000	42.045	0.000	0.088	0.000	1.610	0.000	0.000	0.000	1.000	1.000	1.000
BART	0.690	0.000	0.544	0.000	0.652	0.000	0.612	0.236	0.547	**0.331**	**0.113**	**0.286**
HP	**0.577**	**12.629**	0.504	**0.067**	**0.526**	**1.563**	**0.466**	**0.216**	**0.458**	0.326	0.112	0.284
SP	0.600	−0.798	**0.493**	0.020	0.579	0.525	0.542	0.222	0.487	0.303	0.102	0.266
HP+SP	0.688	−2.034	0.511	0.015	0.643	0.420	0.559	0.237	0.528	0.301	0.099	0.260

Table 5: Results on MAC-Dial. The performance of the model is better when Control Error Rate (CER) is lower ↓, ROUGE is higher ↑, and Control Correlation (CC) is closer to the golden summary ↕.

CER and CC across all 5 control attributes. Compared with the HP model, the SP model has similar control ability on *Ext* and *Spe*. However, it does not perform well on *Len* and *Tpc*. This could be the result of using the pretraining checkpoint that has learned some knowledge about the length-related hard prompt before training (Section 6.3).

Table 5 displays the results of MAC-Dial. Similar to the MAC-Doc dataset, the HP model obtains the highest scores on most of the metrics. However, the overall performance of length decreases because using the pretrained CNNDM checkpoint does not lead to performance gain in the dialogue domain (Section 6.3).

It is worth noting that the CER should not be compared across datasets, because its scale is different from different datasets. For example, random uncontrolled BART in MAC-Doc obtains 1.177 CER for *Ext* while it is 0.544 in MAC-Dial.

5.3 Human Evaluation

Although automatic metrics usually provide a speedy comparison, these metrics cannot easily evaluate the quality of the control, especially mixed-attribute control. Thus, we also conduct a human evaluation for the controlled summaries.

Evaluation Method We hire two evaluators with expertise in English and text summarization. We show them randomly selected summaries generated by different systems with the source text and control attributes. The evaluators answer a yes/no question: ‘‘For the given summary, does it follow the control requirement of this attribute?’’ Specifically, we select golden summaries, summaries generated by HP model, and summaries generated by HP+SP model. For each model, we pick 30 samples from MAC-Doc and MAC-Dial separately, resulting in 180 summaries in total. Furthermore, we compute Cohen’s kappa (Cohen, 1960) to measure the agreement between evaluators.

Evaluation Results Table 6 shows the human evaluation results. Each number (except for kappa) is calculated by the count of yes answers divided by the total count of questions, indicating the control ability of the model. As shown, the HP model performs better than HP+SP on most of the attributes. This result confirms the consistency of our proposed CER and CC with human evaluation.

Besides, golden summaries always rank first, and the kappa score of the two evaluators is

	MAC-Doc				MAC-Dial				
	Tpc	*Ext*	*Spe*	Kappa	*Tpc*	*Spk*	*Ext*	*Spe*	Kappa
Gold	0.83	0.77	0.80	0.87	0.87	0.80	0.73	0.73	0.84
HP	0.67	0.73	0.57	0.77	0.67	0.70	0.67	0.70	0.79
HP+SP	0.53	0.60	0.60	0.70	0.60	0.57	0.53	0.40	0.69

Table 6: Human evaluation results. We evaluate Speaker (*Spk*), Extractiveness (*Ext*), and Specificity (*Spe*). Length does not require human annotation because it is measured by counting the number of tokens.

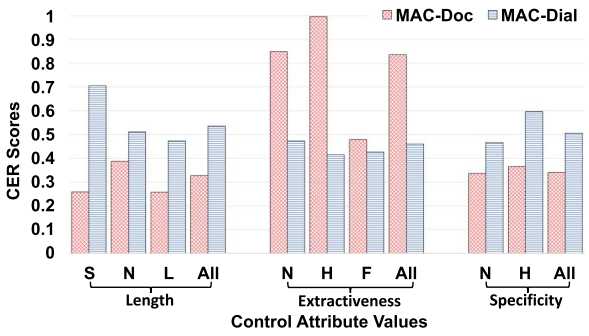
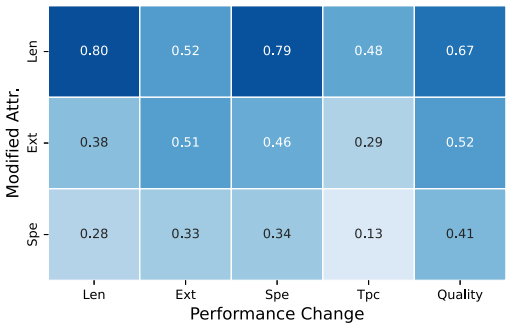


Figure 5: Difficulty of attribute values. The *x*-axis shows the control attribute and its value. For instance, S in length is the CER of all the *Len: short* samples.



(a) MAC-Doc attribute dependency

(b) MAC-Dial attribute dependency

Figure 6: Dependency of attributes. Each row shows the attributes that are modified while each column shows the change in the corresponding attribute.

over 0.8. These two results also verify the high annotation quality of MACSUM, because human evaluators agreed that the golden summaries followed the control requirements most of the time.

6 Analysis and Discussion

For a deeper understanding of the task of mixed-attribute controllable summarization on MACSUM, we conduct analysis including attribute difficulty, attribute dependency, model pretraining, and present several example outputs for case studies.

6.1 Difficulty of Controlling Attribute Values

Models have different difficulties in controlling certain attribute values, as some attribute values can be easier or harder to be controlled. We analyze this by comparing CER for different attribute values of the HP model's outputs. As shown in Figure 5, for MAC-Doc, the system obtains a higher CER on *Len: normal* samples compared with the other two values of *Len*, showing that *normal* is more difficult to control, and the hardest values in controlling *Ext* and *Spe* are both *high*. For MAC-Dial, the hardest values in controlling *Len*, *Ext*, and *Spe* are *short*, *normal*, and *high*, respectively.

6.2 Dependency of Attributes

In mixed-attribute controllable summarization, we notice interesting dependencies among attributes, as changing one attribute influences the other one. To analyze this, we randomly select 200 samples from the test set for each attribute, and randomly change this attribute to another value to form a new sample (e.g., from *Len: long* to *Len: short*). Then, the same HP model, without further training, is used to generate summaries on these new samples. We evaluate the performance difference between the newly predicted summaries $\hat{y}'$ and the originally predicted summaries $\hat{y}$ via CER$(\hat{y}', \hat{y})$.

Figure 6 shows the performance change. As can be seen, for MAC-Doc, *Len* has the highest dependency toward other attributes, while *Spe* has the lowest. For MAC-Dial, *Ext* has the highest dependency, while *Spe* has the lowest. We believe this is because the model in MAC-Doc has a strong

	Len	Ext	Spe	Tpc	Spk	Quality
MAC-Doc	0.315	0.870	0.327	0.254	–	0.776
-CNN	0.361	1.033	0.392	0.346	–	0.777
MAC-Dial	0.454	0.392	0.420	0.373	0.211	0.719
-CNN	0.469	0.422	0.430	0.476	0.201	0.722

Table 7: Ablation on MACSUM on pretraining on CNNDM. MAC-Doc, MAC-Dial denote the HP model initialized with BART-large-cnn, while -CNN uses BART-large checkpoint. Numbers for five control attributes are CER and for Quality are the average of ROUGE-1/2/L.

control ability towards *Len*. Thus, the value change of *Len* will influence more on other attributes.

6.3 Effect of Pretraining

We investigate the effect of pretraining on the control ability of summarization models. For two HP models initialized by BART-large and BART-large-cnn separately, we compare their results after finetuning them on both MAC-Doc and MAC-Dial.

As shown in Table 7, for MAC-Doc, the BART-large-cnn initialized model is able to control the length substantially better than the vanilla BART-large initialized model. On the contrary, for MAC-Dial, the advantage of the BART-large-cnn checkpoint is negligible. Using BART-large-cnn or not only slightly influences the control ability of all attributes in MAC-Dial. We believe the reason for this is that the CNNDM pretraining provides certain useful information for the model to learn the ability to control attributes on news articles.

6.4 Case Study

We show three case studies in Table 8, discussing three typical phenomena in mixed-attribute controllable summarization, namely, Topic Defocus, Length against Specificity, and Extractiveness against Readability.

Topic Defocus In Table 8 Case 1, MACSUM asks for a summary focusing on the topic of ''*education*''. Although the human-annotated summary does not contain the topic word, its contents are still highly related to ''*education*''. This shows that human annotators have the flexibility of conducting high-level summarization of the topic. In contrast, although the model-generated summary contains the topic word, its content is poorly structured. This shows the challenge of topic defocus, a phenomenon where models rely too much on explicitly containing the topic words when generating topic-controlled summaries.

Length against Specificity Another challenge is the contradiction between long length and low specificity. Long summaries contain more tokens and inevitably invite more specific information. On the contrary, short summaries only describe core events using a few words and are naturally biased towards low specificity. As shown in Table 8 Case 2, when *Len* is *short* and *Spe* is *high*, both HP and HP+SP generated summaries are longer compared with the human-annotated summary.

Extractiveness against Readability As shown in Table 8 Case 3, when *Ext* is *full*, the model-generated summaries are choppy and unnatural, in particular for dialogues. When humans are asked to annotate fully extractive summaries, they may have to write unnatural sentences, and this phenomenon is amplified by a trained summarization system. As shown in the table, the HP+SP generated summary is not grammatical and consists of short phrases instead of complete sentences. This can be explained by the fact that the complicated dialogue discourse structures and frequent interactions between different interlocutors make salient information sparse.

7 Conclusion

We propose MACSUM, a high-quality human-annotated benchmark for mixed-attribute controllable summarization. It contains 5 types of control attributes, including Topic, Speaker, Length, Extractiveness, and Specificity. To the best of our knowledge, MACSUM is the first dataset with mixed attributes as well as human annotations. We explore the hard prompt and soft prefix models and evaluate them on MACSUM. Results and analysis demonstrate that hard prompt models yield the best performance and also show this is a challenging task as a large gap between machine learning models and human still exists.

Future work can design more effective models for the mixed-attribute controllable summarization task, or explore mixed-attribute control on other generation tasks.

Case 1: Topic Defocus (MAC-Doc)	
Attributes	Length: normal; Extractiveness: full; Specificity: normal; **Topic: education**
Gold	They quickly reopened the University of Mosul, under a radically altered curriculum. Some subjects would be banned – democracy and political thought, hotel management, tourism and archaeology. ISIS allows girls to go to school, in a segregated environment.
HP	The Taliban, forbids all girls' education. But ISIS allows girls to go to school, albeit in a segregated environment.
Case 2: Length against Specificity (MAC-Doc)	
Attributes	**Length: short**; Extractiveness: normal; **Specificity: high**; Topic: campaign
Gold	All variations of women feature in Sunday's campaign video release by Hillary Clinton.
HP	Clinton's new campaign website is focused on women and the progress of women in politics.
HP+SP	During her campaign, Hillary Clinton has focused on women's issues. Her new campaign website is filled with women's pictures.
Case 3: Extractiveness against Readability (MAC-Dial)	
Attributes	Length: normal; **Extractiveness: full**; Specificity: normal; Topic: parallel marketing, angle; Speaker: Marketing
Gold	Marketing; either market it as the point of view; we could have parallel marketing s schemes; one where you've got one where it appeals to people that want to have the new device that looks cool, is fashionable; So um, I dunno we'll have to decide which which angle we're gonna go to or both; Either market it together by getting control in a set colour or like you buy it with several; as a separate thing.
HP+SP	Marketing, could have parallel marketing, schemes, one where it appeals to people that want to have the new device that looks cool; one that rather, than a kind of a need relationship with the device; people might not like, having a device, just looks nice; also a device, practically sound; decide which angle, gonna go to or both.

Table 8: Three case studies on MACSUM.

Acknowledgments

We thank Ming Zhong, Tao Yu, Haoran Zhang, Sarkar Das, and Nan Zhang for their valuable feedback and suggestions. We also would like to thank the reviewers and action editor for their helpful comments and reviews.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.528

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,

pages 3110–3120, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1307

Jeremy J. Bornstein, Douglass R. Cutting, John D. Hatton, and Daniel E. Rose. 1999. Interactive document summarization. US Patent 5,867,164.

Florian Boudin. 2016. pke: An open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 69–73, Osaka, Japan. The COLING 2016 Organizing Committee.

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*.

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. https://doi.org/10.18653/v1/2021.naacl-main.476

Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained Markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232. https://doi.org/10.1162/tacl_a_00423

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. AdaPrompt: Adaptive model training for prompt-based NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6057–6068, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1046

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. https://doi.org/10.1177/001316446002000104

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*, volume 2005, pages 1–12.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney, Australia. Association for Computational Linguistics.

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. https://doi.org/10.18653/v1/2021.naacl-main.384

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479. https://doi.org/10.1613/jair.1523

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2706

Seeger Fisher and Brian Roark. 2006. Query-focused summarization by supervised sentence

ranking and skewed word distributions. In *Proceedings of the Document Understanding Conference, DUC-2006, New York, USA*. Citeseer.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022a. News summarization and evaluation in the era of GPT-3. *ArXiv preprint*, abs/2209.12356. https://doi.org/10.48550/arXiv.2209.12356

Tanya Goyal, Nazneen Rajani, Wenhao Liu, and Wojciech Kryscinski. 2022b. HydraSum: Disentangling style features in text summarization with multi-decoder models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 464–479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *ArXiv preprint*, abs/2012.04281. https://doi.org/10.48550/arXiv.2012.04281

Pengcheng He, Baolin Peng, Liyang Lu, Song Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. 2022. Z-code++: A pre-trained language model optimized for abstractive summarization. *ArXiv preprint*. https://doi.org/10.48550/arXiv.2208.09770

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.

Ella Hofmann-Coyle, Mayank Kulkarni, Lingjue Xie, Mounica Maddela, and Daniel Preotiuc-Pietro. 2022. Extractive entity-centric summarization as sentence selection using bi-encoders. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.243

Anton Leuski, Chin-Yew Lin, and Eduard Hovy. 2003. iNeATS: Interactive multi-document summarization. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 125–128, Sapporo, Japan. Association for Computational Linguistics. https://doi.org/10.3115/1075178.1075197

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. https://doi.org/10.18653/v1/2020.acl-main.703

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

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.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *ArXiv preprint*, abs/2103.10385. https://doi.org/10.48550/arXiv.2103.10385

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. https://doi.org/10.18653/v1/2022.acl-long.474

Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1444

Annie Louis and Ani Nenkova. 2011. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42, Portland, Oregon. Association for Computational Linguistics.

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. https://doi.org/10.18653/v1/2022.acl-long.237

Takuya Makino, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. 2019. Global optimization under length constraint for neural text summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1099

Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.365

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. https://doi.org/10.1162/tacl_a_00438

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.410

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1044

Giuseppe Russo, Nora Hollenstein, Claudiu Cristian Musat, and Ce Zhang. 2020. Control, generate, augment: A scalable framework for multi-attribute text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 351–366, Online. Association for Computational Linguistics.

https://doi.org/10.18653/v1/2020.findings-emnlp.33

Itsumi Saito, Kyosuke Nishida, Kosuke Nishida, Atsushi Otsuka, Hisako Asano, Junji Tomita, Hiroyuki Shindo, and Yuji Matsumoto. 2020. Length-controllable abstractive summarization by guiding with summary prototype. *ArXiv preprint*, abs/2001.07331. https://doi.org/10.48550/arXiv.2001.07331

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.185

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1099

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020*, pages 8902–8909. AAAI Press.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.510

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wen Xiao, Lesly Miculicich, Yang Liu, Pengcheng He, and Giuseppe Carenini. 2022. Attend to the right context: A plug-and-play module for content-controllable summarization. https://doi.org/10.48550/arXiv.2212.10819

Ming Zhong, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. 2022. Unsupervised summarization with customized granularities. *Findings of EMNLP 2022*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.472

A Implementation Details

We list the implementation details for the models.

HP+SP For HP+SP on MAC-Doc, we load the HP trained model first and set different learning rates for the language model and prefixes, i.e., $3e-5, 1e-6$ separately, and remove the Len prefix from the model. This is because we find that the HP model obtains high performance with Len related attributes very well, due to the pretrained BART-large-CNN checkpoint. Using prefix tuning or tuning the language model with a large learning rate will hurt the performance (Section 6.3). For HP+PE on MAC-Dial, we only set the different learning rates, but we do not load the checkpoint or remove the Len prefix. This is because the CNN pretrained checkpoint is not significantly beneficial for MAC-Dial (Section 6.3).

BART The model is a pretrained BART model which only prepends the hard prompt of topic and speaker to the input, which means it does not control the rest of the attributes. This is the baseline to justify if we control these three attributes.

B Annotator Details

We have four annotators with native English background. Before the pilot test, we also supply annotators with professional training for high-quality annotation and provide annotation visualization tools for the annotators to regularize the annotation process. For each sample, we ask the annotators to inspect the quality and decide to keep the annotation or discard it due to difficulty or errors. We combine the annotations of each week to form the MACSUM dataset with careful processing: we discard the invalid samples reported by the annotators and use a program to filter out the other invalid samples with empty or wrong text.

C Annotation Guidelines

We write annotation guidelines of MACSUM for two purposes. First, the guidelines are used as our criteria to evaluate annotators during the pilot test. Second, during annotation, we provide annotation guidelines to the annotators and ask them to carefully follow them. For both purposes, the guidelines are a key step to ensure the quality of the whole annotation process. Thus, we pick out some of the details in the guidelines. Note that the following paragraphs are directly copied from the guideline document and shared across all four annotators.

Speakers Annotation Criteria. A dialogue may contain multiple speakers. if we specify certain speaker names as the control attribute, it means we only care about what these speakers say in the dialogue. So we need to focus on the dialogue turns spoken by these speakers and write the summary for them.

Topics Annotation Criteria. Topic is represented by a set of keywords (usually) copied from the dialogue. A dialogue may contain multiple topics, we need to summarize the content that is only related to the given topic.

Length Annotation Criteria. *Normal length*: the length of the summary should equal 15%–25% of the related text spans. For example, the dialogue contains 2000 words, and the related text span for the labeled speaker contains 1000 words. Then we need to write 15%–25% x 1000 = 150 – 250 words for the summary. *Long summary*: the length of the summary should equal 30%–35% of the related text spans. *Short summary*: the length of the summary should equal 5%–10% of the related text spans. These criteria should be *dynamically modified*, the target of length control is to differentiate the length of the different outputs. We can adjust the criteria a little bit if the lengths of the three types of summaries are too similar.

Extractiveness Annotation Criteria. *Normal extractiveness*: the same as a natural summary that humans will write. *High extractiveness*: copy more sentences/tokens from the source text compared with normal extractiveness. *Full extractiveness*: copy all the sentences/tokens from the source text. Again, this can be modified if we can better differentiate summaries with different abstractiveness.

Specificity Annotation Criteria. *Normal specificity*: the same as a natural summary that humans will write. *High specificity*: include more descriptive content in the source text compared with normal specificity.

D Examples of the MACSUM Dataset

Table 9 shows five examples of our proposed MAC-Doc dataset. Note that samples 2 and 3, and samples 4 and 5, only differ in *Len*, *Ext*, and *Spe*.

Source text	(CNN)Jackson Gordon is no ordinary 21-year-old. By day he is an industrial design student at Philadelphia University, but Gordon has another side to him – a side altogether darker, tougher, and more enigmatic. Hanging in his workshop Gordon has a full suit of armor plating, cape, and cowl – matte black and built to stop a knife. Gordon has an alter ego: the Dark Knight himself, Batman. You might expect his origin story to be cloaked in mystery, but speaking to CNN Gordon is quick to explain how the transformation took place. . . . Perhaps because of their versatility and the small matter of copyright issues, those that go on sale will not feature the iconic bat symbol. Gordon says his fledgling business will remain small whilst he's at University – he has to finish his studies after all, and won't be using the project towards his degree credits. For now the Batsuit and Armatus Design will remain a one man operation: such is the life of a superhero."
Attributes	Length: short; Extractiveness: normal; Specificity: normal; Topic: (No Topic Specified);
Gold	Jackson Gordon, a 21-year-old industrial design student at Philadelphia University built a Batsuit that is resistant to stabs, knife slashes, and high impacts. According to Gordon, this is a second attempt at building the suit after an earlier attempt five years ago.
HP	The transformation of Jackson Gordon, a 21-year-old industrial design student at Philadelphia University, into a Batman fan has happened. Gordon has created a full suit of armor plating, cape and cowl with suede detailing.
Attributes	Length: normal; Extractiveness: normal; Specificity: normal; Topic: industrial design student;
Gold	Apart from being an industrial design student, Gordon is also a Shaolin Kung Fu expert and has started a business making jackets and cowls but plans to focus on studies first.
HP	The industrial design student Jackson Gordon, 21, is no ordinary student. Gordon has created a replica of the iconic Batman suit with an alter ego named after the Dark Knight.
Attributes	Length: normal; Extractiveness: **fully**; Specificity: **high**; Topic: industrial design student;
Gold	By day he is an industrial design student; Gordon is also an expert in Shaolin Kung Fu; He has already begun manufacturing the cowls for the public.
HP	By day, an industrial design student, Gordon, has another side to him; a side altogether darker, tougher and more enigmatic; Gordon has an alter ego, the Dark Knight himself, Batman; as elaborate as his design was, it lacked the functionality or the authenticity of the genuine article.
Attributes	Length: normal; Extractiveness: normal; Specificity: normal; Topic: conventional materials;
Gold	Gordon chose unconventional materials to build the Batsuit ensuring that every part was protected whether it had armor plates or not.
HP	In order to avoid using conventional materials, Gordon used memory foam, built around key areas to s̈quish and compressäreas to dissipate the impact of blows, also used Kevlar as the base fabric.
Attributes	Length: **long**; Extractiveness: normal; Specificity: normal; Topic: conventional materials;
Gold	Gordon chose unconventional materials, using Kevlar for slash resistance, a form of memory foam for impact absorption, ABS plastic for armor plates, and polyurethane for the cowl.
HP	Eschewing conventional materials, Gordon opted for a form of memory foam, built around key areas to s̈quish and compress,d̈issipating the impact of blows; also used Kevlar as the base fabric, making it cut and slash resistant to bladed weapons, but breathable and wearable all day.

Table 9: Five case studies on MAC-Doc.