CTL-Prompt: Contrastive Topic-Length Prompt Learning for Dialogue Summarization

Anonymous ACL submission

Abstract

The prevalence of online meetings, such as Zoom and Microsoft Teams, has highlighted the necessity for an effective dialogue summary. This study proposes Contrastive Topic-Length Prompt Learning (CTL-Prompt), a simple method that generates topic-based summaries. First, we used topic prompts to direct our dialogue summarization in order to steer the summary towards a particular topic in light of recent success with prompts in guiding aspects in general summarization. Nevertheless, our preliminary experiment revealed that depending solely on the topic prompt frequently leads to mostly identical summaries across topics. We further added a length control prompt that controls the length of the generated summaries based on the length of the reference summaries for each topic. While it was able to generate a more concise summary, the summaries across topics remained similar. To promote the model to produce concise yet diverse summaries across topics, we propose the use of contrastive learning on topic-length prompts, which make use of positive and negative pairs to enforce the models to learn the similarities and differences of different topics. Experimental results showed that our model outperformed other baseline models in the ROUGE and BERT scores on the DialogSum dataset. This result was reproduced in the MACSum dataset, and similar results were found. Our work is available at [anonymized].

1 Introduction

Dialogue summarization condenses key information from a dialogue and presents it in a more concise form, enabling individuals to quickly grasp the essential points. A lot of different ideas were put forward, such as using pre-trained summarization models (Khalifa et al., 2021; Chen et al., 2021; Feng et al., 2021), graph-based methods to understand complex relationships (Zhao et al., 2020a; Chen and Yang, 2021), multi-encoders to understand different points of view in dialogues (Chen and Yang, 2020), contrastive learning to understand when people talk about similar topics at the same time (Tang et al., 2021; Liu et al., 2021), and more.

Dialogue Example	
#Person1#:	Are you enjoying your trip to New Orleans?
#Person2#:	Oh, yes. I really like it here.
#Person1#:	Would you like to do something tonight?
#Person2#:	Sure. I'd love to.
#Person1#:	Let's see. Have you been to a jazz club yet?
#Person2#:	Yes. I've already been to several clubs here.
#Person1#:	OK. What about an evening riverboat tour?
#Person2#:	Uh, actually, I've gone twice this week.
#Person1#:	So, what do you want to do?
#Person2#:	Well, I haven't been to the theater in a long time.
#Person1#:	Oh, OK. I hear there's a terrific show at the Sanger Theater.
#Person2#:	Great! Let's make a reservation.
Gold Summary1:	#Person1# and #Person2# are discussing where to have fun, and they decide to go to the theater tonight.
Gold Summary2:	#Person1# and #Person2# are talking about what to do tonight and they finally decide to go to watch a show.
Gold Summary3:	#Person2# hasn't been to the theater for a long time, so #Person1# and #Person2# decide to make a reservation for a show at the Sanger Theater.
BART$_{large}$:	#Person1# invites #Person2# to a jazz club, an evening riverboat tour, and a show at the Sanger Theater.
T 1:	#Person2# enjoys the trip to New Orleans. #Person1# suggests an evening riverboat tour and a show at the Sanger Theater.
T 2:	#Person2# enjoys the trip to New Orleans. #Person1# suggests an evening riverboat tour and a show at the Sanger Theater.
T 3:	#Person2# enjoys the trip to New Orleans. #Person1# suggests an evening riverboat tour and a show at the Sanger Theater.
T-L 1:	#Person1# invites #Person2# to a jazz club and an evening riverboat tour in New Orleans tonight.
T-L 2:	#Person1# invites #Person2# to a jazz club and an evening riverboat tour in New Orleans. They finally decide on a terrific show.
T-L 3:	#Person1# invites #Person2# to a jazz club and an evening riverboat tour in New Orleans, and they finally decide to go to the Sanger Theater.
T-L-CL (Ours) 1:	#Person1# invites #Person2# to a jazz club or an evening riverboat tour in New Orleans. #Person2# chooses the theater.
T-L-CL (Ours) 2:	#Person1# invites #Person2# to a jazz club, an evening riverboat tour, and a terrific show at the Sanger Theater.
T-L-CL (Ours) 3:	#Person1# invites #Person2# to go to a jazz club or an evening riverboat tour in New Orleans. #Person3# suggests going to the Sanger Theater.

Figure 1: A typical pretrained model such as BART$_{large}$ produces a generic single summary. Topic prompts (T) generate mostly identical summaries across topics. Conciseness can be enhanced by using the topic prompt paired with the length prompt (T-L), but it remained producing similar summaries across topics. Our proposed technique (T-L-CL) generates concise yet diverse summaries relevant to the specified topic. (Note: Topic 1: *"Leisure activity"*; Topic 2: *"Terrific show"*; Topic 3: *"Theater"*; Note 2: Text color signifies longest common summaries across topics; Note 3: Five more samples are provided in Appendix.)

However, despite the advancements, relatively less work has been done on topic-guided dialogue summarization. This is relevant accounting for the fact that a dialogue involves multiple speakers with different perspectives, intents, and actions (Chen et al., 2021). In other words, it can be beneficial to allow users to generate a summary that is relevant to their interests. Indeed, there were few such attempts, e.g., Amplayo et al. (2021) allowed users to control opinion summaries by specifying aspects; similarly, Xu and Lapata (2020) proposed query-focused summarization for multi-document summarization, which summarizes multiple documents based on a given query. In any case, such a technique usually requires modification of model architectures or requires a query as input for training. In addition, little work has been done specifically on topic-guided dialogue summarization. In recent years, the idea of prompting has attracted much interest due to its simplicity (e.g., it does not require the modification of the model architecture). For example, Zhang et al. (2022) achieved controllable summarization through prompts that use control signals (e.g., length of generated summaries, named entities that appear in summaries) during the model training phase. Nevertheless, the use of prompts remains underexplored in the area of dialogue summarization.

This study proposes Contrastive Topic-Length Prompt Learning, a simple method that generates topic-based summaries. We chose DialogSum (Chen et al., 2021) as it closely represents real-world situations. First, we used topic prompts to guide our dialogue summarization. Nevertheless, our preliminary experiment revealed that depending solely on the topic prompt frequently leads to mostly identical summaries across topics (see Figure 1 and see more in the Appendix.). We further add a length control prompt, as introduced in Wang et al. (2022). Nevertheless, while it helps in producing more concise summaries, the generated summaries remain similar across topics. Inspired by the recent interest in contrastive learning, we apply contrastive learning to the topic-length prompt, which was found to help produce concise yet diverse summaries across different topics. Specifically, it makes use of positive and negative pairs, which helps enforce the model to better distinguish different topics. We found that contrastive learning is especially useful to learn multiple topics during the training phase, even when topic annotation is limited. For example, in the DialogSum, only a single topic summary is available for the training set, while the testing set contains three topic summaries, which resemble real-world cases of scarce topic annotations.

Our experimental results showed that our model outperformed other baseline models in the ROUGE and BERT scores. We also further experimented with variations of negative samples. The contributions to our work are as follows:

1. Our proposal involves the utilization of Contrastive Topic-Length Prompt Learning for the purpose of dialogue summary.

2. Our simple prompt-based method achieved superior performance compared to the baseline models on the DialogSum and MACSum datasets.

3. We have conducted experiments and analyses comparing prompt variants and variants of negative samples, yielding few research insights.

2 Related Work

2.1 Dialogue Summarization

It's hard to summarize dialogue because there are many people involved, the subject changes, there are a lot of cross-references, there are different kinds of interaction cues, and the language is specific to the conversation (Feng et al., 2021). The generation of a dialogue summary still faces issues with repetition, a lack of variation, incoherence, and lack of topic-guided summarization (Sun and Li, 2021).

BART (Lewis et al., 2019) is an encoder-decoder model that has been widely employed in dialogue summarization. Khalifa et al. (2021) discovered that BART performed better than UniLM and other conventional abstractive methods when tested on the SAMSum (Gliwa et al., 2019) dataset. On the DialogSum dataset, which is highly abstractive and resembles real-life scenarios, Chen et al. (2021) found that BART performance on DialogSum is similar to that used by the UniLM model.

Graph-based techniques were presented to address the intricate relationships in dialogue summarization. Zhao et al. (2020b) proposed a graph-attention-based mechanism to encode long-distance relationships within the dialogue. Chen and Yang (2021) utilized a structured graph to

model "who does what" to input to the graph attention network for better dialogue summarization. However, the use of a graph-based technique is often not suitable for parallelization and can be computationally demanding.

2.2 Guided Summarization

Guided summarization can be performed by directly modifying the model architecture or using a prompt.

2.2.1 Modifying Architectures

Amplayo et al. (2021) proposed the use of aspect controllers, which pool the tokens, sentences, and documents that are most relevant to the user's specified aspect. Xu and Lapata (2022) proposed query-focused summarization for multi-document summarization. Both works excel in enabling users to input certain aspects or queries that can direct the summary process. However, these models require a query as input during training.

Other approaches have also been proposed. Chen and Yang (2020) proposed a multi-view decoder model that takes in hidden states from multiple encoders that encode different views, and the decoder decides the attention weights on which view it should focus on to produce the final summary. Zhong et al. (2022) proposed a pre-trained methodology using masking techniques for dialogue summarization. In any case, both works do not consider topic-guided summarization.

Specifically, in the area of topic-guided dialogue summarization, Liang et al. (2023) proposed a global-local centrality model to help select the salient context from all sub-topics. Here, the global one aims to identify vital sub-topics in the dialogue, and the local one aims to select the most important context in each sub-topic. Finally, it is used to guide the model to capture both salient context and sub-topics when generating. In addition, Zou et al. (2021) proposed a topic-oriented summarization model for customer service dialogues. Specifically, it is a topic-augmented two-stage dialogue summarizer for a customer, jointly with a saliency-aware neural topic model.

However, it is worth noting that a potential disadvantage is that these works necessitate altering the architecture of the model.

2.2.2 Prompt-based Approaches

Recently, there has been a growing interest in the use of prompts to regulate summarization, mostly because of their simplicity. They have the ability to manipulate the characteristics of produced summaries and potentially enhance the quality of summarising. For example, Zhang et al. (2022) achieved controllable summarization through prompts. They used control signals (e.g., length of generated summaries, named entities that appear in summaries) during the model training phase. Wang et al. (2022) introduced a simple prompt design that specifically control length of generated summaries. On the other hand, Zhang et al. (2023) included speaker, topic, length, specificity and extractiveness as prompt to control the summary generation (but found that only topic and speaker were useful). Based on its simplicity, we begin to explore the possible combination of a prompt-based approach with contrastive learning for topic-guided dialogue summarization.

2.3 Contrastive Learning

Contrastive learning was proposed as a means to gain a deeper understanding of facets discussed in a dialogue. CONFIT (Tang et al., 2021) incorporated contrastive loss to mitigate the issues of missing information and incorrect references in conversation summarization tasks. Xiong et al. (2023) utilized contrastive learning as a means to decrease repetition in the context of scientific summarization. Tan and Sun (2023) has shown that using contrastive learning may improve the quality of the output summary by letting the model tell the difference between training falls caused by false negative samples. Liu et al. (2021) make use of contrastive learning by forcing the models to contrast positive and negative samples, where positive samples are defined based on a specified window utterance size, allowing the decoder to capture salient intent information. Regardless, none of this work specifically focuses on the production of topic-guided summarization. The success of these techniques motivates us to use contrastive learning for the purpose of topic-guided dialogue summarization.

3 Methodology

We propose contrastive topic prompt learning, a method that enables summary generation based on a specified topic. Here, we chose DialogSum (Chen et al., 2021) as it closely represents real-world situations. Specifically, DialogSum comprises a triple of document, topic and summary $\{(D, T, S)\}$ where a document is coupled with a topic and a summary

in the training set, while in testing, a document is coupled with a set of topics $T = \{t_1, t_2, t_3\}$ along with its respective summaries $S = \{s_1, s_2, s_3\}$.

As aforementioned, merely using a topic-length prompt is inadequate, as it often leads to identical summaries across different topics. To enlarge the distance between different topics, we make use of positive and negative topic examples. Specifically, the actual topic (i.e., specified by the dataset) serves as a positive, and its synonyms and random topic words serve as negative samples to force the models to learn the similarities and differences of summaries in different topics. Note that although it is more intuitive to use synonyms as positive samples, due to the very small difference between synonyms and the actual topics, synonyms are promising candidates to serve as *hard* negative samples, similar to the discussion of *hard negative mining* discussed in Robinson et al. (2020).

Finally, given the input, the objective is to minimize two losses namely, the contrastive loss and the negative log-likelihood to generate output summary.

3.1 Prompt Template

Here we introduce our prompt template that guides the generation using topic. Specifically, we frame our input as `Topic of Summary:`$\{t\}$`, Dialogue:` $\{d\}$ where t denotes the topic and d is our dialogue context. To train our model using contrastive learning, the topic t serves as a positive sample (t_p) and its synonym and random topic word serve as negative samples (t_n). Here, the ratio between synonyms and random topics is kept equal. Note that it is important to consider that while using synonyms as positive samples may seem more intuitive, synonyms can actually be effective as challenging negative samples due to their close resemblance to the actual topics. This concept is similar to the idea of hard negative samples discussed in Robinson et al. (2020) work on contrastive learning.

As for the synonym replacement, we employ wordnet of the NLTK library. For the given topic t, we obtain a set of synonyms $Syn_t = \{syn_1, syn_2, ..., syn_{|Syn_t|}\}$ in which we randomly select one as the replacement of the topic and represent as negative sample. On the other hand, for random word topics, we randomly select a topic word given in the training dataset.

Note that for the length control, we additionally included `Length of Summary:`$\{l\}$ as a part of our prompt template. Here, l denote the length of a summary used during the training phase which is simply a number of summary words defined by space (i.e., `string.split`).

Hence, our final prompt template becomes, `Topic of Summary:` $\{t\}$`. Length of Summary` $\{l\}$`. Dialogue:` $\{d\}$, where t denote topic, l denote length and d is our dialogue context.

3.2 Contrastive Learning

Our framework incorporated contrastive learning to assist the learning. In particular, it makes use of positive and negative pairs to enforce the model and learn the similarities and differences of summaries on different topics. Specifically, we obtained the last hidden state of the encoder of positive and negative topic prompts and employed the typical max-margin contrastive loss function as follows:

$$\mathcal{L}_{con} = \max(0, \cos(h_1, h_2) - \text{margin}) \quad (1)$$

where h_1 denote the last hidden state of positive samples, while h_2 denote the last hidden state of negative samples and the margin is set to 0.5.

3.3 Dialogue Summarization

To generate dialogue summary, we perform fine-tuning on the pretrained model. Given the input, the objective is to minimize a joint loss namely the contrastive learning and the cross entropy losses of generating the output summary $s = \{s_1, s_2, ..., s_{|s|}\}$. The cross entropy loss is defined as negative log-likelihood (NLL) as follows:

$$\mathcal{L}_{nll} = -\sum_{i=1}^{|s|} f\left(s_i | D, s_{<i}\right) \quad (2)$$

where $f\left(s_i | D, s_{<i}\right)$ is the log-likelihood of the ith token of the reference summary.

Hence, total loss becomes,

$$\mathcal{L}_{total} = \mathcal{L}_{nll} + \alpha \mathcal{L}_{con} \quad (3)$$

Where $\mathcal{L}_{nll}$ is negative log-likelihood (NLL) and $\mathcal{L}_{con}$ is contrastive learning loss, it integrates NLL with loss by multiplying with alpha, which defaults to 0.5.

4 Experiments

4.1 Datasets

In DialogSum, a training sample comprises a document coupled with a topic and a summary. To clarify, DialogSum provides only one topic summary

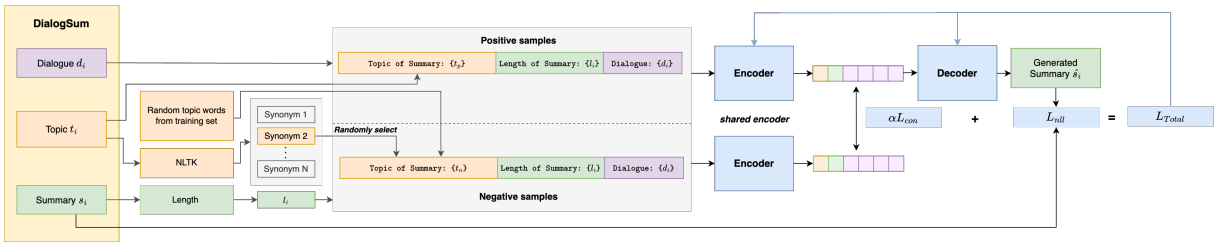

Figure 2: Overview of our framework. Here our prompt template is constructed by framing the input as `Topic of Summary:`$\{t\}$`, Dialogue:` $\{d\}$ where t denote topic and d is our dialogue context. To train our model using contrastive learning, the topic t serves as positive (t_p) and its synonym and random topic word serve as negative samples (t_n). The samples are then passed to the model with the objective as to minimize two losses namely, the contrastive loss and the negative log-likelihood to generate output summary.

per document in the training set. This limitation in topic annotation makes DialogSum challenging and resembles real-world scenarios (Chauhan et al., 2022). In testing set, a document is coupled with a set of three topics and their respective summaries. Specifically, the dataset is collected from various sources, including DailyDialog, DREAM, and MuTual, and consists of 13,460 daily conversations, which are divided into three subsets: 12,460 for training, 500 for validation, and 1500 for testing. Note that 1500 is derived from an initial set of 500 samples, and each of these samples addresses three distinct topics.

4.2 Experimental Setting

Here, we describe the experimental setting of our work. Our implementation is based on the BART$_{large}$ model, which contains 406M parameters. Here, all input was truncated to 1024, and the output is set to 128 tokens. For the fine-tuning, the learning rate is set to 5e-05, and the model was trained for 15 epochs at batch size 4 with min and max output lengths of 1 and 128, respectively. Additionally, we adopt AdamW as our optimizer and gradient accumulation is set to 32. At inference time, a beam size of 4 is selected, with the min and max output lengths kept the same as fine-tuning. The experiment was run on one A6000 GPU.

As for the evaluation metric, we used three types of ROUGE score, which is the main metric in almost all summarization tasks. ROUGE-1 measures the overlap of unigrams. ROUGE-2 measures the overlap of bigrams. ROUGE-L measures the longest common sub-sequence between a candidate summary and a reference summary. In addition, the BERT score (Zhang et al., 2019) was also used to understand semantic comparisons between generated and reference summaries.

5 Results

We experimented (1) the prompt design which includes the comparison with two baselines - pretrained BART$_{large}$ (Lewis et al., 2019) and the current SOTA for DialogSum, i.e., LA-BART (Wang et al., 2022) - with topic prompt (`T`), with topic prompt + length control (`T-L`), with topic prompt + contrastive learning (`T-CL`), and with topic prompt + length control + contrastive learning (`T-L-CL`), (2) negative samples selection for contrastive learning, where we experimented using random words, synonyms, and combined in an equal ratio as negative samples. Note that random words here refer to words randomly picked from the list of topic words in the training set.

5.1 Prompt Design

Table 1 shows the comparison between different prompt designs and the baselines.

Our four topic-prompt based designs outperformed the `Baseline (LA-BART-LARGE)` and `Baseline (BART-LARGE)` in most scores. `T-L` and `T-L-CL` were among the best performer in most scores.

To understand whether the summaries were identical across the topics, we calculated the number of longest n-gram normalized by length between combination of three generated summaries. Results showed that `T-L-CL` outperformed other variants, suggesting that `T-L-CL` was able to generate diverse summaries across topics.

5.2 Contrastive Learning

We designed an experiment to explore the use of synonyms and random words as our negative samples to assist contrastive learning. Note that random words are words randomly picked from the list of topic words in the training set. Specifically, we

Prompt	R-1			R-2			R-L			BERTScore	N-gram	Len. Δ
	P	R	F1	P	R	F1	P	R	F1			
Baseline (BART-LARGE)	44.55	53.26	47.10	19.94	23.46	20.87	42.51	49.31	44.72	0.9183	0.990	6.97
Baseline (LA-BART-LARGE)	48.03	50.89	48.95	21.73	22.86	22.07	45.84	47.95	46.56	0.9216	0.660	3.56
Ours (T)	44.30	54.53	47.39	19.89	23.85	20.98	42.22	50.15	44.85	0.9180	0.642	7.84
Ours (T-CL)	43.31	**54.89**	46.99	19.25	23.87	20.61	41.36	**50.31**	44.45	0.9175	0.622	8.17
Ours (T-L)	48.98	52.33	**50.22**	22.62	**23.97**	**23.09**	46.65	49.16	47.62	0.9229	0.538	3.22
Ours (T-L-CL)	**50.36**	50.73	50.10	**23.17**	23.19	22.97	**47.92**	48.08	**47.70**	**0.9230**	**0.529**	**2.95**

Table 1: Comparison of different prompt designs in DialogSum. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L recall respectively. Len.Δ refers to the difference in the number of tokens between the generated and the reference summary (i.e., whether the generated summaries are overly long or short). N-gram scores refer to the average number of longest n-grams normalized by length between the three generated summaries. The highest scores are bolded. Here the performance of our designs are compared against two baselines - pretrained BART$_{large}$ (Lewis et al., 2019) and the current SOTA for DialogSum, i.e., LA-BART. The designs include topic prompt (`T`), topic prompt + length control (`T-L`), topic prompt + contrastive learning.(`T-CL`), and topic prompt + length control + contrastive learning (`T-L-CL`).

Prompt	Positive	Negative	R-1			R-2			R-L			BERTScore	Len. Δ
			P	R	F1	P	R	F1	P	R	F1		
T-L-CL	Actual Topic	Random	47.97	52.47	49.53	22.12	23.99	22.73	45.71	49.10	46.96	0.9216	4.34
T-L-CL	Actual Topic	Synonym	47.93	**53.08**	49.60	21.95	**24.02**	22.57	45.65	**49.44**	46.96	0.9216	5.50
T-L-CL	Actual Topic	Random, Synonym	**50.36**	50.73	**50.10**	**23.17**	23.19	**22.97**	**47.92**	48.08	**47.70**	**0.9230**	**2.95**

Table 2: BERTScore and delta length Precision, Recall and F1-score in ROUGE metric and BERT score on three types of negative samples. Here the performance of our proposed method (T-L-CL) using both random topic words and synonym as negative samples is compared against one with random topic words only and synonym only as negative samples to assist contrastive learning.

compared synonyms alone, random topic words alone, and a combination of both in an equal ratio to our negative samples. Table 2 shows that using a combination of both yielded the highest results in terms of F1 scores, while using random topic words alone yielded the highest recall scores and synonyms alone yielded the highest precision scores.

6 Discussion

We present a discussion on several interesting details of our findings.

6.1 Prompt Designs

Looking more deeply, `T` and `T-CL` were the common best performers in recall scores. This can be linked to the non-conciseness of their summaries. Note that recall is high when the generated summary contains all the words in the reference summary, but the drawback could be its non-conciseness. Thus, longer-generated summaries tend to have a high recall. To further understand this, we calculated the Len.Δ, which was measured by the difference between the number of tokens in the generated summary and the reference summary. We found that `T` and `T-CL` scored the highest Len.Δ, which suggested that the high recall score could be from the overly long generated summary.

On the other hand, `T-L` was able to constrain the length for more concise summaries, as seen in the better precision. As for its recall, it is expected to achieve a slightly lower score due to its shorter length. In any case, `T-L` performed worse than `T-L-CL` in the number of longest n-gram scores, as well as all precision scores.

Lastly, our personal experience identified the clear tradeoffs between recall and precision. In `T` and `T-CL` conditions, though the recall score was the highest, we had difficulty increasing the precision score, leading to non-concise summaries. In `T-L`, we were able to effectively increase the precision score (i.e., summary becoming more concise), but at the same time, we also observed lower recall scores. The interesting aspect we found was that contrastive learning was an effective method that allowed us to maintain both recall and precision.

6.2 Contrastive Learning

Our results showed that a combination of synonyms and random topic words achieved the best performance in terms of F1 scores. On the other hand, using random topic words alone achieved scores very similar to the `T-L` condition (as if no contrastive learning was used).

A potential explanation is to look at the con-

Prompt	R-1			R-2			R-L			BERTScore	N-gram	Len. Δ
	P	R	F1	P	R	F1	P	R	F1			
Baseline (BART-LARGE-CNN)	32.17	32.84	30.01	10.19	9.50	9.16	27.02	27.78	25.82	0.8551	1.00	34.29
Baseline (LA-BART-LARGE-CNN)	29.26	36.03	29.63	9.44	10.76	9.23	24.60	29.72	25.22	0.8529	0.924	40.49
T-S (MACSum Paper)	41.08	36.02	34.94	16.70	14.40	14.06	35.66	31.92	31.42	0.8684	0.345	34.22
Ours (T-S-CL)	**42.67**	38.03	36.47	**17.92**	15.90	**15.27**	**36.92**	33.33	**32.46**	**0.8699**	0.328	35.22
Ours (T-S-L)	40.93	39.00	36.12	16.83	15.81	14.77	35.40	33.92	32.11	0.8681	**0.273**	37.34
Ours (T-S-L-CL)	40.78	**39.24**	**36.55**	16.79	**15.92**	14.98	35.20	**34.26**	**32.46**	0.8693	0.303	**34.40**

Table 3: Comparison of different prompt designs in MACSum. Extra configuration includes `S` which refers to speaker prompt. Our experiment found that speaker prompt is consistently useful for MACSum thus we hold this condition constant for all conditions.

trastive loss equation. When random topic words were used, the cosine similarity of the two words may be low, hence the max margin loss may result in near 0, rendering the contrastive loss null. On the other hand, when synonyms were used, the cosine similarity of the two words were relatively high, hence the max margin loss becomes high, thus forcing the model to differ the generated summaries. When compare these three generated summaries with the test set which contains the three different topic summaries, precision got increased.

As to why using both synonyms and random topic words achieved the best, this matches to the original paper regarding hard negative sampling (Robinson et al., 2020). It empirically found that overly big β (which control the amount of negative samples) is not necessarily good, since overly large β strongly prefers pushing hard negative samples away for which other soft, easier negative samples are not accounted for, resulting in a overly tight, non-generalized boundary.

6.3 MACSum Dataset

To better understand how contrastive learning contributes when the nature of the dataset changes, we implemented our technique on another dialogue summarization dataset, specifically the MACSum dataset. One notable difference is that MACSum contains an average reference summary length of 69.4 tokens, while DialogSum only has an average summary length of 18.8 tokens. Another notable difference is that the MACSum training set contains as many as 10+ topic summaries. Thus, using MACSum allowed us to determine whether contrastive learning remains effective when the nature of the dataset changes.

A brief explanation of the dataset is as follows. The MACSum dataset is a human-annotated dataset that bears resemblance to the DialogSum dataset. MACSum specifically integrates source texts from two separate domains, news stories and dialogues with human annotations. These annotations include information such as length, extractiveness, specificity, topic, and speaker. MACSum is separated into three subsets: 2338 for training, 292 for validation, and 324 for testing. Full experimental settings can be found in the Appendix.

Table 3 shows the results and the Appendix shows some examples of the generated summaries. First, all our prompt designs outperformed the baselines in most scores which also include the original MACSum paper. Comparing conditions with `CL` and its non-CL variants, it is clear that contrastive learning did indeed help improve performance, as seen in the increased performance from `T-S` to `T-S-CL`, and from `T-S-L` to `T-S-L-CL`. The enhanced performance is more evident in MACSum, as compared to DialogSum, possibly as a result of the greater number of topic summaries accessible in the MACSum training set; for instance, one dialogue can contain up to ten topic summaries in MACSum, thereby facilitating the contrastive learning process even more effectively.

It is important to see that `T-S-L` did better than `T-S-L-CL` in terms of n-gram scores. The very plausible reason for this is that the DialogSum training set contains only one topic per dialogue, resembling a real-world situation of limited topic annotations. Consequently, contrastive learning (CL) aids in comprehending the distinctions between topics, resulting in more varied summaries. In MACSum, the training set includes numerous topic summaries. Therefore, even without `CL`, `T-S-L` was able to identify the distinctions between topics and generate a variety of summaries based on the specified topics. The higher ROUGE ratings though shows that `CL` still contributes to producing more aligned summaries that are in line with the given topics. Overall, CL proved to be effective, irrespective of the characteristics of the dataset.

One noteworthy observation is the relatively diminished influence of `L` in comparison to its effect

in DialogSum. A key observation is that MACSum has an average reference summary length of 69.4 tokens, but DialogSum only has an average summary length of 18.8 tokens. In addition, it is important to mention that MACSum contains a diverse reference summary lengths, ranging from 10 tokens to as much as 400 token. Therefore, it is plausible that a basic length prompt may not sufficiently convey to the model the desired level of conciseness for the summary, given the significant deviations in length among summaries. One potential avenue for future research could involve utilizing different prompt designs that can help generate very long summaries.

7 Conclusion and Future Work

We propose Contrastive Topic-Length Prompt Learning, a simple yet effective method that generates topic-based summaries. Specifically, to guide the summary towards a specific topic, a topic-length prompt is utilized. Additionally, we propose contrastive learning on prompts, which allows the model to generate less identical yet concise summaries on different topics. The experimental results showed that our model outperformed baseline models in ROUGE scores on the DialogSum and MACSum datasets. Future work includes the inclusion of more exhaustive explorations of contrastive learning techniques and loss functions, numbers and types of negative samples, and prompt variations.

8 Acknowledgement

The authors would like to thank the [anonymized].

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.

Vipul Chauhan, Prasenjeet Roy, Lipika Dey, and Tushar Goel. 2022. Tcs_witm_2022@ dialogsum: Topic oriented summarization using transformer based encoder decoder model. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 104–109.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. *arXiv preprint arXiv:2010.01672*.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. *arXiv preprint arXiv:2109.08232*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xinnian Liang, Shuangzhi Wu, Chenhao Cui, Jiaqi Bai, Chao Bian, and Zhoujun Li. 2023. Enhancing dialogue summarization with topic-aware global-and local-level centrality. *arXiv preprint arXiv:2301.12376*.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. *arXiv preprint arXiv:2109.04994*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Shichao Sun and Wenjie Li. 2021. Alleviating exposure bias via contrastive learning for abstractive text summarization. *arXiv preprint arXiv:2108.11846*.

Caidong Tan and Xiao Sun. 2023. Colrp: A contrastive learning abstractive text summarization method with rouge penalty. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *arXiv preprint arXiv:2112.08713*.

Bin Wang, Chen Zhang, Chengwei Wei, and Haizhou Li. 2022. A focused study on sequence length for dialogue summarization. *arXiv preprint arXiv:2209.11910*.

Jing-Wen Xiong, Xian-Ling Mao, Yizhe Yang, and Heyan Huang. 2023. Cplr-sfs: Contrastive prompt learning to reduce redundancy for scientific faceted summarization. In *Journal of Physics: Conference Series*, volume 2506, page 012006. IOP Publishing.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*, pages 3632–3645.

Yumo Xu and Mirella Lapata. 2022. Document summarization with latent queries. *Transactions of the Association for Computational Linguistics*, 10:623–638.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yubo Zhang, Xingxing Zhang, Xun Wang, Si-qing Chen, and Furu Wei. 2022. Latent prompt tuning for text summarization. *arXiv preprint arXiv:2211.01837*.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. Macsum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020a. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020b. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11765–11773.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14665–14673.

A MACSum Experimental Setting

Here, we describe the experimental setting of our experiments on MACSum dataset. MACSum comprises two subcategories; MAC-Doc and MAC-Dial. Specifically, we focus on MAC-Dial which was collected from QM-Sum. Our implementation is based on the $\text{BART}_{largecnn}$ model, which has 406M parameters. Here, all input was truncated to 1024, and the output is set to 400 tokens. For the fine-tuning, the learning rate is set to 3e-05, and the model was trained for 30 epochs at batch size 6 with min and max output lengths of 1 and 400, respectively. Additionally, we adopt AdamW as our optimizer and gradient accumulation is set to 32. At inference time, a beam size of 4 is selected, with the min and max output lengths kept the same as fine-tuning. The experiment was run on one A100 GPU.

B MACSum Prompt Template

Here we introduce our prompt template that guides the generation for MACSum dataset. We used the topic to do contrastive learning, similar to how we did on DialogSum. Furthermore, we include the speaker as an extra attribute following the topic prompt, as described in (Zhang et al., 2023). They utilized both the speaker and topic as prompts for the model's input. To confirm, our preliminary experiment also found that without the speaker prompt, it consistently performed more poorly across all conditions thus we include it in all our prompt designs. Note that MACSum also incorporates extractiveness and specificity features that we currently do not utilize.

Hence, our final prompt template becomes, `Topic of Summary:` $\{t\}$. `Speaker` $\{s\}$. `Length of Summary` $\{l\}$. `Dialogue:` $\{d\}$, where t denote topic, s denote speaker, l denote length and d is our dialogue context.

C Examples

Here we provided five DialogSum examples. Due to MACSum long sumamries, we provided only two MACSum dialogue samples and their generated summaries.

DialogSum Example 1	
#Person1#:	I need to use the ATM.
#Person2#:	What's stopping you?
#Person1#:	I'm not sure how.
#Person2#:	I don't understand. It is pretty easy.
#Person1#:	I've never used one before.
#Person2#:	OK. I can help you figure it out.
#Person1#:	What do I have to do?
#Person2#:	First, slide your card into the machine.
#Person1#:	Then what?
#Person2#:	You need to type your PIN in.
#Person1#:	What do I have to do next?
#Person2#:	Click on whichever option you want, and you're done.
#Person1#:	Thank you!
Gold Summary1:	#Person1# doesn't know how to use the ATM. #Person2# teaches #Person1# step by step.
Gold Summary2:	#Person1# doesn't know how to use an ATM. #Person2# teaches #Person1#.
Gold Summary3:	#Person2# is teaching #Person1# how to use an ATM step by step because #Person1# doesn't know how to use.
BART$_{large}$:	#Person2# teaches #Person1# how to use the ATM.
T 1:	#Person2# teaches #Person1# how to use the ATM.
T 2:	#Person2# teaches #Person1# how to use the ATM.
T 3:	#Person2# teaches #Person1# how to use the ATM.
T-L 1:	#Person2# teaches #Person1# how to use the ATM, which is a new experience for Person1.
T-L 2:	#Person2# teaches #Person1# how to use the ATM.
T-L 3:	#Person2# teaches #Person1# how to use the ATM and tells Person 1 it's pretty easy. #Person2# #Person3# tells the steps in detail.
T-L-CL (Ours) 1:	#Person2# teaches #Person1# how to use the ATM, which is very surprising.
T-L-CL (Ours) 2:	#Person2# shows #Person1# how to use the ATM.
T-L-CL (Ours) 3:	#Person1# needs to use the ATM but doesn't know how. #Person2# shows how to do it.

Figure 3: Appendix: Sample generated summaries. Note: Topic 1 - "*use the ATM*"; Topic 2 - "*ATM*"; Topic 3 - "*ATM using*".

DialogSum Example 2	
#Person1#:	Mr. Blake? Mr. Foster's on the phone. He'd like to know if you can send over those training manuals?
#Person2#:	Oh, tell him I'll leave them at his office tomorrow afternoon.
#Person1#:	He was hoping that you could drop them off this afternoon.?
#Person2#:	I'm afraid that I can't do that. They're at the printer's being copied. They'll be back tomorrow before 1 o'clock.
Gold Summary1:	Mr. Blake explains the training manuals cannot be sent today because they are still being copied.
Gold Summary2:	#Person1# is transferring the message between Mr. Blake and Mr. Foster about the training manuals.
Gold Summary3:	Mr. Foster wants the training manuals to be sent this afternoon but Mr. Blake explains that they haven't been printed out yet.
BART$_{large}$:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're at the printer being copied.
T 1:	Mr. Blake tells #Person1# he can't send over the training manuals because they're at the printer.
T 2:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're being copied.
T 3:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're being copied.
T-L 1:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're at the printer.
T-L 2:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're being copied.
T-L 3:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're at the printer's being copied.
T-L-CL (Ours) 1:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're being copied.
T-L-CL (Ours) 2:	Person1 tells Mr. Blake Mr. Foster wants him to send over the training manuals.
T-L-CL (Ours) 3:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon because they're at the printer.

Figure 4: Appendix: Sample generated summaries. Note: Topic 1 - "*office phone*"; Topic 2 - "*training manuals*"; Topic 3 - "*send training manuals*".

DialogSum Example 3	
#Person1#:	Well, I'm glad I had a chance to see these samples, but I'm really not in a position to make a decision right now.
#Person2#:	Would you like me to leave the samples with you?
#Person1#:	Yeah, why don't you do that, and then you can give me a call sometime next week.
#Person2#:	Fine, Mr. Grant. Thanks for your time. I'll talk to you next week.
Gold Summary1:	#Person2# leaves the samples to Mr. Grant and will talk to him next week.
Gold Summary2:	#Person2# leaves Mr. Grant with samples and will call Mr. Grant later to know the decision.
Gold Summary3:	#Person2# leaves the samples at Mr. Grant's so he can make a decision with more consideration.
BART$_{large}$:	Mr. Grant asks #Person2# to leave the samples with him.
T 1:	Mr. Grant asks #Person2# to leave the samples with him and will give him a call next week.
T 2:	Mr. Grant asks #Person2# to leave the samples with him and will give him a call next week.
T 3:	Mr. Grant asks #Person2# to leave the samples with him and will give #Person1# a call next week.
T-L 1:	Mr. Grant asks #Person2# to leave the samples with him and will call next week.
T-L 2:	Mr. Grant asks #Person2# to leave the samples with him and will give him a call next week.
T-L 3:	Mr. Grant asks #Person2# to leave the samples with him and will give him a call next week.
T-L-CL (Ours) 1:	Mr. Grant asks #Person2# to leave the samples with him and will call next week.
T-L-CL (Ours) 2:	Mr. Grant asks #Person2# to leave the samples with him and will make a decision next week.
T-L-CL (Ours) 3:	Mr. Grant tells #Person2# he's not in a position to make a decision now.

Figure 5: Appendix: Sample generated summaries. Note: Topic 1 - "*sample*"; Topic 2 - "*transaction*"; Topic 3 - "*office conversation*".

MACSum Example 1	
Project Manager :	'Kay . Alright . Now we have Courtney with the functional requirements .
Marketing :	Yes , okay so we tested a hundred subjects in our lab , and we just we watched them and we also made them fill out a questionnaire , and we found that the {vocalsound} users are not typically happy with current remote controls . Seventy five percent think they're ugly . Eighty percent want {disfmarker} they've {disfmarker} are willing to spend more , which is good news for us um if we make it look fancier , and basically w we just need something that really I mean there's some other points up there , but they {disfmarker} it needs to be snazzy and it {disfmarker} but yet simple .
User Interface :	gap Wait .
Marketing :	So that's really what we need to do . And we need we need it to be simple , yet it needs to be high-tech looking . So {disfmarker}
User Interface :	And that meaning what ?
Marketing :	Like {disfmarker} They like I guess use the buttons a lot .
.	.
.	.
.	.
Project Manager :	{vocalsound} Didn't they {disfmarker} um didn't our rival companies manufacture a remote that you would press the button on the TV and it would {disfmarker} the remote would beep so if you have lost it {disfmarker}
User Interface :	It's kinda like what the remote phone used to do .
Project Manager :	Mm . Oh , yeah , that's true .
User Interface :	You know like go to the base .
Project Manager :	We could definitely include that if we wanted to .
User Interface :	Yeah .
Project Manager :	If it's within our price . Okay . Are we ready for our last presentation , Amber ?
Gold Summary1:	Marketing said that they tested hundreds of subjects in a study, about remotes, seventy-five percent think it was ugly, eighty percent wanted to spend money if they make it look fancier. Something simple, thirty-four percent said that learning to use a new remote was hard. The most used buttons were the volume, the power buttons, and the channel buttons. It said speech recognition could be something coming on.
Gold Summary2:	Marketing; tested a hundred subjects in our lab; watched them and we also made them fill out a questionnaire; found that the users are not typically happy with current remote controls; spend more; make it look fancier; User Interface; only use ten percent of the buttons; Marketing; need something simple, because most people, well thirty four percent say that it's just too much time to learn how to use a new one; don't want to vary too far from the normal standard remote.
Gold Summary3:	Marketing; it needs to be snazzy; but yet simple; because most people, well thirty four percent say that it's just too much time to learn how to use a new one; don't want to vary too far from the normal standard remote; need to be able to identify it, and RSI; Project Manager; It's very important; User Interface; ergonomics like your hand movements or something; Marketing; the channel, the volume and the power buttons are the most important on our company website; need a modern eye-catching design; really needs to be simple; need a small number of buttons; also need to take in like are most people going to be using it for a DVD player, a TiVo; need a hip, but not a corny marketing scheme for promoting our product; findings found that people are willing to pay more money for a remote for it to be more high-class we could consider it.
BART$_{large}$:	Marketing said that users were not typically happy with current remote controls and that they needed something simple and snazzy but yet high-tech looking. User Interface added that users only used ten percent of the buttons.
T 1:	Marketing said that they tested a hundred subjects in their lab, and they found that the users were not typically happy with current remote controls, and that they needed something that was snazzy and simple, but high-tech looking.
T 2:	Marketing said that they tested a hundred subjects in their lab, and they found that users were not typically happy with current remote controls, that they thought they were ugly, and that they were willing to spend more if they made it look fancier, so they needed something that was snazzy and simple, but high-tech looking, with some other points up there, but they only used ten percent of the buttons, and users didn't really need all the buttons that were contained on there.
T 3:	Marketing said that users were not typically happy with current remote controls, and that they needed to make it look fancier, snazzy, and simple, with some other points up there, but it needed to be high-tech looking.
T-L 1:	Marketing; tested a hundred subjects in our lab; found that the users are not typically happy with current remote controls; Seventy five percent think they're ugly; Eighty percent want; they've been willing to spend more; if we make it look fancier; just need something that really I mean there's some other points up there; need to be snazzy and it needs to be; but yet simple; need it to be simple; high-tech looking; use the buttons a lot.
T-L 2:	Marketing said that they tested a hundred subjects in their lab, and they found that the users were not typically happy with current remote controls. Seventy-five percent thought they were ugly. Eighty percent were willing to spend more if they made it look fancier. Marketing said that users didn't really need all of the buttons that were contained on there, because they only used ten percent of the button really.
T-L 3:	Marketing said that users were not typically happy with current remote controls, and that they needed to make it look fancier and snazzy, and simple, but high-tech looking. Marketing also said that it needed to be simple, because most people said it would take too much time to learn how to use a new one.
T-L-CL (Ours) 1:	Marketing said that they tested a hundred subjects in their lab, and they found that the users were not typically happy with current remote controls, and that they needed something that was snazzy, but yet simple. And they needed it to be simple, yet high-tech looking.
T-L-CL (Ours) 2:	Marketing said that they tested a hundred subjects in their lab, and they found that the users were not typically happy with current remote controls. Seventy-five percent thought they were ugly. Eighty percent were willing to spend more if they made it look fancier. Marketing said that users didn't really need all of the buttons that were contained on there, because they only used ten percent of them.
T-L-CL (Ours) 3:	Marketing said that users were not typically happy with current remote controls, and that they needed to make it look fancier and snazzy, with some other points up there, but it needed to be simple and high-tech looking.

Figure 6: Appendix: Sample generated summaries. Note: Topic 1 - "*functional requirements*"; Topic 2 - "*design*"; Topic 3 - "*remote*".

MACSum Example 2	
Project Manager :	alright ? Great . Um , alright , and I'm sure that , um um , the glow-in-the-dark , fluorescent , whatever , system , um is a go ahead . Is everyone interested in that ?
Industrial Designer :	Y
Marketing :	On the buttons ?
User Interface :	I I like the light up suggestion . I think that would be better .
Industrial Designer :	Yeah .
Project Manager :	Yeah .
User Interface :	'Cause you know the way fluorescent lights lose their brightness after certain time , so
Industrial Designer :	Yeah . Yeah .
Project Manager :	Yeah .
Industrial Designer :	Yeah .
Marketing :	{gap} it doesn't {disfmarker}
User Interface :	I would go for {disfmarker}
Marketing :	It could it could be a tactile thing as well . Um right , if w if we're minimising buttons , we might be able to make them actually larger . And there's something on it . S you know like up arrow down arrow for , for volume .
Industrial Designer :	Like a raised {disfmarker}
Marketing :	Um , and I don't know what we could do for , for channels . S
User Interface :	Well just the numbers could be embossed , couldn't it ? Like raised .
Marketing :	The numbers themselves .
User Interface :	Yeah . Could be raised .
Marketing :	But then the like up button and down button for the channel , channel changing .
User Interface :	Just little arrows , that you could feel , maybe ?
Industrial Designer :	Yeah {disfmarker}
.	.
.	.
.	.
Project Manager :	Alright , well . Let's have lunch and we'll discuss this later .
User Interface :	{vocalsound} Okay .
Marketing :	{vocalsound}
Project Manager :	Alright ?
Industrial Designer :	Okay . Thank you .
Gold Summary1:	User Interface said that fluorescent lights lost their brightness after a certain time and the fact that the presence or absence of the glow-in-the-dark or light-up made no difference to the visually impaired, also that numbers usually wore off if they were just painted or printed on the buttons. More tactile features like bigger buttons, embossed numbers, and arrows, or buttons in the shape of numbers and arrows, were suggested to be incorporated to make the remote available to everybody. Also, the material suggested for the buttons would be softer rubber material instead of hard buttons in order to cater to people with repetitive strain injury.
Gold Summary2:	The product could be personalized by incorporating different colors of interchangeable faceplates into the design. The plates could be made to advertise certain TV programs. However, the idea of advertising certain TV programs had the potential of causing copyright problems
Gold Summary3:	Marketing raised concerns about the proposed lighting up feature sucking up a lot of battery power, suggesting that the battery thing could be tackled by incorporating self-timed lights that came on upon touching any button and stayed on for a few seconds only.
BART$_{large}$:	Project Manager asked if they were interested in the glow-in-the-dark, fluorescent, or fluorescent, system. Marketing thought that it might be sucking more battery power if it was a light up. Project Manager said that it wouldn't make any difference anyway. Marketing said that big buttons for people who were visually impaired might be better.
T 1:	Project Manager asked if everyone was interested in the glow-in-the-dark, fluorescent, whatever, system. User Interface said that it would be better because fluorescent lights lose their brightness after certain time. Marketing said it could be a tactile thing as well, if they minimised the buttons, they might be able to make them larger.
T 2:	Project Manager asked if the glow-in-the-dark, fluorescent, whatever, system was a go ahead. Marketing said that it might be sucking more battery power if it was a light up. Industrial Designer said that big buttons for people who were visually impaired wouldn't make any difference anyway.
T 3:	Marketing thought that the glow-in-the-dark or light up might be sucking more battery power.
T-L 1:	Project Manager asked if everyone was interested in the glow-in-the-dark, fluorescent, system, on the buttons. User Interface liked the light up suggestion and suggested that it could be a tactile thing as well. Industrial Designer said that big buttons for people who were visually impaired wouldn't make any difference and that the buttons could be in the shape of the numbers themselves and be made out of some glow in the dark material.
T-L 2:	Project Manager said that the glow-in-the-dark, fluorescent, system, was a go ahead. Marketing asked about the buttons that would light up. User Interface said that it would be better if it was a tactile thing. Industrial Designer said that big buttons for people who were visually impaired wouldn't make any difference.
T-L 3:	Marketing thought that the glow-in-the-dark or light up might be sucking more battery power, if there, if it is a light up.
T-L-CL (Ours) 1:	User Interface liked the light up suggestion and User Interface suggested it could be a tactile thing as well because if they were minimising buttons, they might be able to make them actually larger and there was something on it like up arrow down arrow for volume. User Interface also suggested that just the numbers could be embossed.
T-L-CL (Ours) 2:	User Interface liked the light up suggestion and suggested it could be a tactile thing as well. User Interface suggested the numbers could be embossed and the numbers themselves could be made out of some glow-in-the-dark material. Project Manager suggested incorporating them both so that the buttons could be in the shape of the numbers.
T-L-CL (Ours) 3:	Marketing said that it might be sucking more battery power, if it was a light up, so they could incorporate them both so that the buttons could be in the shape of the numbers themselves and be made out of some glow-in-the-dark material.

Figure 7: Appendix: Sample generated summaries. Note: Topic 1 - "*fluorescent buttons*"; Topic 2 - "*personalization*"; Topic 3 - "*battery thing*".