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

Anonymous ACL submission

Abstract

The prevalence of online meeting has highlighted the necessity of dialogue summary. Topic summarization is one domain attracted much interest from industry. Anyhow, past work either use topic- or length-prompt which tend to generate almost identical summaries across similar and even different topics. This study proposes Contrastive Topic-Length Prompt Learning (CTL-Prompt), a simple method that generates topic-based summaries. To produce concise yet diverse summaries across topics, we propose contrastive learning on topic-length prompts, which leverages positive and negative pairs to allow the models to learn the similarities and differences of topics. Results showed that our model outperformed baseline models in the ROUGE, BERTscores, and human evaluation scores on the DialogSum and the MACSum dataset. Our work can be found at [anonymized].

1 Introduction

Dialogue summarization condenses key information from a dialogue into a concise form. Different ideas were put forward, such as using pre-trained summarization models (Khalifa et al., 2021; Chen et al., 2021; Feng et al., 2021a), graph-based methods (Zhao et al., 2020a; Chen and Yang, 2021), multi-encoders to understand different perspectives in dialogues (Chen and Yang, 2020), contrastive learning to understand when people talk about similar topics (Tang et al., 2021; Liu et al., 2021).

However, relatively less work has been done on topic-guided dialogue summarization. It can be beneficial to allow users to generate a summary that is relevant to their interests. A few such attempts were presented, e.g., Amplayo et al. (2021) allowed users to control opinion summaries by specifying aspects; Xu and Lapata (2020) proposed query-focused summarization for multi-document summarization. In any case, such a technique usually requires modification of model architectures. In recent years, the idea of prompting has attracted much interest due to its simplicity. For example, Zhang et al. (2022a) achieved controllable summarization through prompts that use control signals

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 do something tonight. They decide to go to the Sanger Theater and make a reservation.
T-L-CL (Ours) 2:	#Person2# enjoys the trip to New Orleans and #Person1# suggests going to a jazz club, an evening riverboat tour, and a terrific show.
T-L-CL (Ours) 3:	#Person2# enjoys the trip to New Orleans. #Person1# suggests a jazz club, an evening riverboat tour, and a show at 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. Despite pairing topic prompt with the length prompt (T-L), the summaries remained similar across topics. Our proposed technique (T-L-CL) generates different summary (T-L-CL 1) and diverse summaries (T-L-CL 2 & T-L-CL 3) 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.)

(e.g., length of generated summaries, named entities that appear in summaries) during the training phase. Nevertheless, our preliminary work demonstrated that merely using prompt is insufficient.

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 but our preliminary experiment revealed that depending solely on the topic prompt frequently leads to mostly identical summaries across topics (see Figure 1 and more in the Appendix.). We further add a length control prompt, as introduced in Wang et al. (2022a). While it helps in producing more concise summaries, the generated summaries remain similar across topics. Inspired by the utility in contrastive learning, we propose contrastive learning to the topic-length prompt, which was found to help produce concise yet diverse summaries across topics. Specifically, we found that contrastive learning remains robust 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, BERTscores, and human evaluation scores.

The contributions are as follows:

1. We propose Contrastive Topic-Length Prompt Learning.

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

3. We have conducted experiments and analyses, yielding further research insights.

2 Related Work

We reviewed related areas of research: (1) dialogue summarization, and (2) guided summarization.

2.1 Dialogue Summarization

Dialogue summarization is a challenging task. This is because many people are involved, the subject changes, many cross-references, involve interaction cues (Feng et al., 2021a). Hence, dialogue summary generation still faces issues with repetition, a lack of variation, incoherence, and lack of topic-guided summarization (Sun and Li, 2021).

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

Note that none of these works focused on topic guided summarization.

2.2 Guided Summarization

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

2.2.1 Modifying Architectures

Amplayo et al. (2021) proposed aspect controllers which pool the tokens, sentences, and documents relevant to user's specified aspect. Xu and Lapata (2022) proposed query-focused summarization for multi-document summarization. Liang et al. (2023) proposed a global-local centrality model to help select the salient context from all sub-topics. Zou et al. (2021) proposed a topic-oriented summarization model for customer service dialogues using a topic-augmented two-stage dialogue summarizer. However, it is worth noting that these works necessitate altering the architecture of the model.

2.2.2 Prompt-based Approaches

There has been a growing interest in the use of prompts for summarization due to its simplicity. Zhang et al. (2022a) used control signals (e.g., length of generated summaries, named entities that appear in summaries) during the model training phase. Yoo and Lee (2023) and You and Ko (2023) extracted keywords from dialogue and utilized them together with prompts to guide the summary generation. Wang et al. (2022b) generated the predicate-argument spans of the dialogue, and utilize them to guide summary generation. Wang et al.

(2022a) utilized control length prompt for summaries generation. 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 explored the possible exploitation of a prompt-based approach, for topic-guided summarization, which we found that merely prompt is insufficient.

2.3 Contrastive Learning

Contrastive learning was proposed to understand facets discussed in a dialogue. CONFIT (Tang et al., 2021) incorporated contrastive loss to mitigate the issues of missing information and incorrect references in dialogue summarization. Xiong et al. (2023) utilized contrastive learning to decrease repetition in scientific summarization. Tan and Sun (2023) found that contrastive learning improve abstractive summarization. Liu et al. (2021) proposed 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 focuses on topic-guided summarization.

3 Methodology

We propose contrastive topic prompt learning. 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 $S = \{s_1, s_2, s_3\}$.

Merely using a topic-length prompt is inadequate, as it often leads to identical summaries across topics. We make use of positive and negative topic examples. In particular, the actual topic (i.e., specified by the dataset) serves as positives, while the negative is defined as similar and random topics. Here, similar topics are those that are similar to the actual topic based on cosine similarity measures and random topics are selected at random from the pool. By doing so forces the models to learn the similarities and differences of topics. Hence, the model is encouraged to generate diverse summaries when given similar topics and generate different summaries when given different topics.

Note that it may seem more intuitive to use similar topics as positive samples, but similar topics 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.

3.1 Prompt Template

We frame our input as `Topic of Summary:`$\{t\}$, `Dialogue:` $\{d\}$ where t denotes the topic and d is the dialogue context. To train the model using contrastive learning, the topic t serves as a positive sample (t_p) and its similar and random topic word serve as negative samples (t_n). The ratio between similar and random topics was experimented.

To obtain the similar topic, for the actual topic t, we first compute the similarity scores between the embedding of dialogue and topics given in the training set using cosine similarity measure:

$$sim_{doc,t_i} = \text{sim}(e_{dial}, e_{t_i}) \quad (1)$$

where e_{dial} denote the embedding of the dialogue, while e_{t_i} denote the embedding of the topic i^{th} from the training set. These scores are then ranked and a set of candidate topics $C_t = \{c_1, c_2, ..., c_{|Ct|}\}$, are selected. Doing so ensures that the candidate topics are related to the dialogue. Next, we compute the similarity scores between the embedding of the actual topic and the candidate topics. This is to obtain the most similar topic for that particular dialogue. It is defined as follows:

$$sim_{actual,c_j} = \text{sim}(e_{actual}, e_{c_j}) \quad (2)$$

where e_{actual} denote the embedding of the actual topic, while e_{c_j} denote the embedding of the candidate topic j^{th} of the dialogue. Note that all the embeddings are obtained by performing an inference on pretrained encoder model.

For random word topics, we randomly select a topic word 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 split by space (i.e., `string.split`).

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

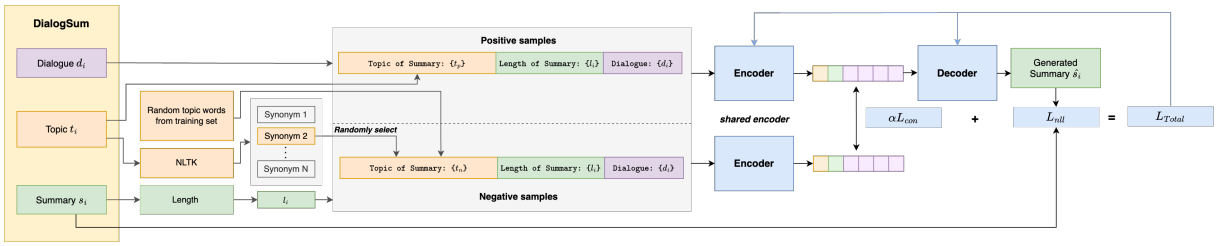

Figure 2: Overview of our framework. Here our prompt template is constructed by framing the input as `Topic of Summary:`$\{t\}$`, Dialogue:` $\{d\}$ where t denote topic and d is our dialogue context. To train our model using contrastive learning, the topic t serves as positive (t_p) and its similar 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.

3.2 Contrastive Learning

Our contrastive learningmakes use of positive and negative pairs to allow the model and learn the diversity and difference of summaries on similar and 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:

$$\mathcal{L}_{con} = \sum_{n \in \{similar, random\}} \max(0, \cos(h_p, h_n) - \text{margin}_n) \tag{3}$$

where h_p denote the last hidden state of positive samples, while h_n denote the last hidden state of negative samples. Here the margin is set to be the mean value of similarity scores between last hidden states of positive and negative samples (see Appendix A)

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:

$$\mathcal{L}_{nll} = -\sum_{i=1}^{|s|} f\left(s_i | D, s_{<i}\right) \tag{4}$$

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,

$$\mathcal{L}_{total} = \mathcal{L}_{nll} + \alpha \mathcal{L}_{con} \tag{5}$$

Where $\mathcal{L}_{nll}$ is negative log-likelihood and $\mathcal{L}_{con}$ is contrastive learning loss; alpha was set to 0.5.

4 Experiments

4.1 Datasets

In DialogSum, a training sample comprises a dialogue coupled with a topic and a summary. DialogSum provides only one topic summary per dialogue 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 dialogue 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 - 12,460 for training, 500 for validation, and 1500 for testing.

4.2 Experimental Setting

For the summarization model, the implementation is based on the BART$_{large}$ model, which contains 406M parameters. 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. We adopt AdamW and gradient accumulation is set to 32. At inference time, a beam size of 4 is selected. The experiment was run on a A6000 GPU.

For the similarity score, 'distilbert-base-nli-mean-tokens' SentenceTransformer was employed. All input was truncated to 128 and mean-pooling was performed.

For the evaluation metric, we used three types of ROUGE score. 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 BERTscore (Zhang et al., 2019) and human evaluation scores were also used.

5 Results

We experimented (1) the prompt design which includes the comparison with two baselines - pre-trained BART$_{large}$ (Lewis et al., 2019) and the current SOTA for DialogSum, i.e., LA-BART (Wang et al., 2022a) - 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 topics, similar topics, and combined in an equal ratio as negative samples.

5.1 Prompt Design

Table 1 shows the comparison between different prompt designs and the baselines. Our four topic-prompt based designs outperformed the `LA-BART-large (baseline)` and `BART-large (baseline)` 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 explore the use of similar and random topics as negative samples. Table 2 shows that using a combination of both yielded the highest results in terms of F1 scores, while using random topics alone yielded the highest precision scores and similar topics alone yielded the highest recall scores.

6 Discussion and Analyses

We discuss accordingly and present further analyses.

6.1 Recall vs. Precision

`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 confirmed that `T` and `T-CL` scored the highest Len.Δ, which suggested that the high recall score came 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, we 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

While using similar topics alone and random topics alone showed strong results, the combination of both achieved the best performance in terms of F1 scores. One potential explanation to why the combination achieved best performance is to look deeper into the focused dataset which is DialogSum in our case. Though DialogSum contains three topics in the test set, they have similar/same meanings, and yet have diverse corresponding summaries. For example, given a dialogue, the topics for summaries are "public transportation", "transportation" and "discuss transport" and given another, the topics are "greeting", "a short visit" and "farewell". Here we notice that some dialogues contain topics that are similar to each other, while some contain relatively different topics. Hence, combining both similar and random topics enforces the model to learn the similarities and differences of summaries on similar and different topics to generate diverse summaries. Here the margins for DialogSum are found to be 0.6 and 0.5 for similar and random topics, respectively. Note that the margins could be adjusted according to one's use.

6.3 Cosine Similarity

To examine the effect of contrastive learning further, we obtained the embedding of the input from

Prompt	R-1			R-2			R-L			BERTScore	N-gram	Len. Δ
	P	R	F1	P	R	F1	P	R	F1			
BART-large (baseline)	44.55	53.26	47.10	19.94	23.46	20.87	42.51	49.31	44.72	0.9183	0.990	6.97
LA-BART-large (baseline) (Wang et al., 2022a)	48.03	50.89	48.95	21.73	22.86	22.07	45.84	47.95	46.56	0.9216	0.660	3.56
KADS (Yoo and Lee, 2023)	-	-	45.99	-	-	20.94	-	-	38.17	-	-	-
TIDSum (You and Ko, 2023)	-	-	48.02	-	-	21.80	-	-	46.15	-	-	-
Fact-aware RL (Wang et al., 2022b)	-	-	48.76	-	-	22.34	-	-	45.29	-	-	-
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)	44.97	53.84	47.60	20.42	23.93	21.35	42.81	49.73	45.07	0.9186	0.656	7.21
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)	**50.10**	51.54	**50.38**	**22.95**	23.38	22.96	**47.52**	48.52	**47.73**	**0.9232**	**0.533**	**3.04**

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	**50.77**	49.73	49.81	**23.10**	22.43	22.56	**48.00**	47.01	47.21	0.9229	**2.55**
T-L-CL	Actual Topic	Similar	48.88	**52.06**	50.07	22.11	**23.42**	22.58	46.38	**48.75**	47.30	0.9229	3.10
T-L-CL	Actual Topic	Random, Similar	50.10	51.54	**50.38**	22.95	23.38	**22.96**	47.52	48.52	**47.73**	**0.9232**	3.04

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 similar topics as negative samples is compared against one with random topic words only and similar topics only as negative samples to assist contrastive learning.

Prompt	R-1			R-2			R-L			BERTScore	N-gram	Len. Δ
	P	R	F1	P	R	F1	P	R	F1			
BART-large-cnn (baseline)	32.17	32.84	30.01	10.19	9.50	9.16	27.02	27.78	25.82	0.8551	1.00	34.29
LA-BART-large-cnn (baseline)	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 (Zhang et al., 2022b)	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)	41.71	**40.27**	37.02	**17.78**	**16.80**	**15.60**	36.17	**35.00**	32.90	0.8710	0.312	38.98
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)	**42.72**	39.29	**37.27**	17.62	16.35	15.50	**36.79**	34.22	**33.04**	**0.8722**	0.291	**36.05**

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. Note that margin of contrastive learning is 0.5 for both similar and random topics.

the last encoder layer from `BART-large`, `T-S-L` and `T-S-L-CL`. The results show that contrastive learning widens the distance (lower cosine similarity) between inputs of different topics to encourage the model to generate diverse and different summaries (see Figure 3).

6.4 MACSum Dataset

We cross-checked our technique on 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

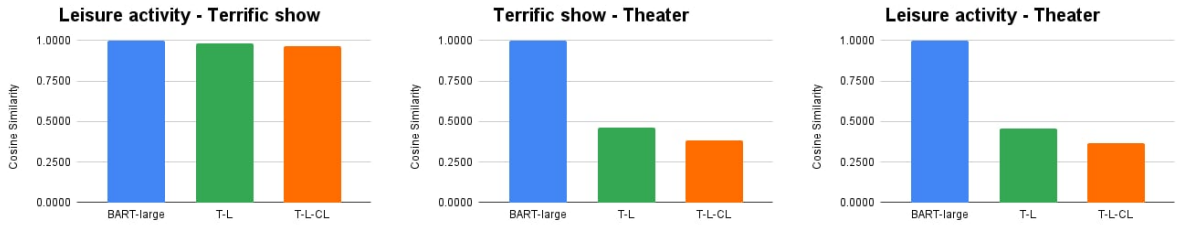

Figure 3: The cosine similarity between two inputs of different topics.

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, contrastive learning 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.

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 tokens. 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.

Model	DialogSum			
	Info.	Conc.	Cov.	Rel.
Gold summary	3.24	3.59	3.12	4.23
BART-large	2.92	3.03	2.93	4.02
LA-BART-large	2.83	3.17	2.96	3.99
Ours (T-L-CL)	3.10	3.38	3.11	4.14

Table 4: Human evaluation results on DialogSum. "Info." is short for informativeness, "Conc." for conciseness, "Cov." for coverage and "Rel." for relevancy.

7 Human Evaluation

Following Feng et al. (2021b), we conducted human evaluation on three metrics for the qualitative measure i.e. informativeness (Inf.), Conciseness (Con.), Coverage (Cov.). Specifically, informativeness evaluates how well the generated summaries capture more salient information. Conciseness measures how well the summary discards redundant information and Coverage measrues how well the summary covers each part of the dialogue. Additionally, we also include Relevancy as one of our metrics. Relevancy measures how well the summary is relevant to the topic.

Specifically, we randomly sampled 10 dialogues with corresponding three topics and generated summaries from both DialogSum and MACSum to conduct the evaluation. Note that for MACSum since one dialogue may contain more than three topics, we randomly select any three and their corresponding summaries for the evaluation. To reduce variance caused by humans, we have 5 human evaluators in which they were asked to rate each summary on the scale of 1 to 5 (higher is better) for each metric. The results are shown in Table 4 - 5. Results showed that our method achieved higher scores than both baselines across all metrics in both DialogSum and MACSum.

Model	MACSum			
	Info.	Conc.	Cov.	Rel.
Gold summary	3.19	3.43	3.33	4.34
BART-large	2.38	2.29	2.33	2.66
LA-BART-large-cnn	2.11	2.00	2.08	2.16
Ours (T-S-L-CL)	3.00	3.52	3.23	4.23

Table 5: Human evaluation results on MACSum. "Info." is short for informativeness, "Conc." for conciseness, "Cov." for coverage and "Rel." for relevancy.

8 Conclusion

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.

9 Limitations

In this study, we applied our proposed method on DialogSum and MACSum datasets, both of which provide dialogue-topic-summary triples. Particularly, we make use of the given topics in the training set to find the similar and random topics for our contrastive learning. In addition, it is also important to note that the margin for the negative samples is also specific to the focused dataset. Hence, this could impact the hinder the generalizability of our method.

10 Acknowledgement

We extend our sincere appreciation to [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. 2021a. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021b. Language model as an annotator: Exploring dialogpt for dialogue summarization. *arXiv preprint arXiv:2105.12544*.

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. 2022a. A focused study on sequence length for dialogue summarization. *arXiv preprint arXiv:2209.11910*.

Ye Wang, Xiaojun Wan, and Zhiping Cai. 2022b. Guiding abstractive dialogue summarization with content planning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3408–3413.

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.

Chongjae Yoo and Hwanhee Lee. 2023. Improving abstractive dialogue summarization using keyword extraction. *Applied Sciences*, 13(17):9771.

Jaeah You and Youngjoong Ko. 2023. Topic-informed dialogue summarization using topic distribution and prompt-based modeling. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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. 2022a. 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. 2022b. Macsum: Controllable summarization with mixed attributes. *arXiv preprint arXiv:2211.05041*.

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.

	DialogSum		MACSum	
	Similar	Random	Similar	Random
mean	0.6	0.5	0.5	0.5
std	0.2	0.1	0.1	0.1
25%	0.5	0.4	0.4	0.4
50%	0.6	0.4	0.5	0.5
75%	0.8	0.5	0.6	0.6

Table 6: Mean value of similarity scores between last hidden states of positive and negative samples. These values are set as margins of negative samples.

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.

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 Margins of DialogSum and MACSum

Here we provided the margin values for the contrastive loss. Note that the values are rounded to one decimal place.

B 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.

C 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.

D 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 #Person1# it's pretty easy. #Person2# #Person3# tells the steps in detail.
T-L-CL (Ours) 1:	#Person1# wants to use the ATM but doesn't know how. #Person2# instructs #Person1#.
T-L-CL (Ours) 2:	#Person2# teaches Person1 how to use the ATM.
T-L-CL (Ours) 3:	#Person1# wants to use the ATM but doesn't know how. #Person2# teaches #Person1# how to use it.

Figure 4: 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:	#Person1# asks Mr. Blake to send over the training manuals but he can't do so today.
T-L-CL (Ours) 2:	Mr. Blake tells #Person1# he can't send over the training manuals this afternoon.
T-L-CL (Ours) 3:	#Person1# asks Mr. Blake to send over the training manuals to Mr. Foster tomorrow afternoon, but Mr.Blake says he can't do that because they're at the printer.

Figure 5: Appendix: Sample generated summaries. Note: Topic 1 - "*office phone*"; Topic 2 - "*training manuals*"; Topic 3 - "*send training manuals*".

DialogSum Example 3	
#Person1#:	What shall we do now?
#Person2#:	Let's get the table wares we need first, And then stand in the line.
#Person1#:	I've got all I need.
#Person2#:	Let's stand in the line now.
#Person1#:	It smells good. Look at these. They make my mouth watering.
#Person2#:	Me too.
#Person1#:	It will be our turn soon.
#Person2#:	Tell the waiter what you want when it's your turn.
#Person1#:	OK. I'll have many cream cakes today.
#Person2#:	I don't like them. I think they are too icky."
Gold Summary1:	#Person1# and #Person2# are waiting for food.
Gold Summary2:	#Person1# and #Person2# are standing in line to buy food.
Gold Summary3:	#Person1# and #Person2# are waiting in line for food.
BART$_{large}$:	#Person1# and #Person2# get the table wares and stand in the line to order food.
T 1:	#Person1# and #Person2# get the table wares and stand in the line for food.
T 2:	#Person1# and #Person2# get the table wares and stand in the line in a restaurant.
T 3:	#Person1# and #Person2# get the table wares and stand in the line to order.
T-L 1:	#Person1# and #Person2# are waiting for food.
T-L 2:	#Person1# and #Person2# stand in the line in a restaurant.
T-L 3:	#Person1# and #Person2# are waiting for the waiter.
T-L-CL (Ours) 1:	#Person1# and #Person2# are waiting for food.
T-L-CL (Ours) 2:	#Person1# and #Person2# get the table wares and stand in the line.
T-L-CL (Ours) 3:	#Person1# and #Person2# are waiting for their order.

Figure 6: Appendix: Sample generated summaries. Note: Topic 1 - "*waiting for food*"; Topic 2 - "*in a restaurant*"; Topic 3 - "*wait for order*".

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-S 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-S 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-S 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-S-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-S-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-S-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-S-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. Seventy-five percent thought they were ugly. Eighty percent want to spend more, which was good news for them if they made it look fancier. Marketing said they just needed something that was snazzy and it needed to be high-tech looking.
T-S-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, and 80 percent were willing to spend more if they made it look fancier, so they needed something that was snazzy and simple, but high-tech looking. Marketing also 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-S-L-CL (Ours) 3:	Marketing said 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 to make it snazzy and high-tech, yet simple.

Figure 7: Appendix: Sample generated summaries. Note: Topic 1 - "*functional requirements*"; Topic 2 - "*design*"; Topic 3 - "*remote*".

MACSum Example 2	
Project Manager :	Okay . Okay , let's talk about all of our {disfmarker} We'll come to decision later about all the components that we need to include , let's um wrap up this one , and {vocalsound} I'm gonna go back to my PowerPoint , 'cause we need to discuss the new project requirements which you might've already seen flashed up on the screen a bit earlier . {vocalsound} Wait , come back . Alright . Sorry , let's go through this . Alright . Here we go . New product requirements . First it's only going to be a TV remote . We're trying not to over-complicate things . So no DVD , no TiVo , no stereo .
Industrial Designer :	{vocalsound} Okay .
Project Manager :	It's not gonna be multi-functional .
User Interface :	{gap}
Project Manager :	Hey . And we th need to promote our company more , so we need to somehow include our colour and our company slogan on the remote . We're trying to get our name out there in the world . Okay .
User Interface :	{gap}
Project Manager :	And you know what teletext is ?
User Interface :	Yeah .
Project Manager :	{gap} in States we don't have it , but um it's like they just have this channel where just has news and weather , kind of sports , User Interface : I know .
Marketing :	What is it ?
User Interface :	{vocalsound}
Project Manager :	it's very um bland looking , it's just text on the screen ,
User Interface :	Yeah ,
Project Manager :	not even {disfmarker}
User Interface :	it's like black , black and white kind of {disfmarker}
Project Manager :	Yeah , just black with just text .
Marketing :	Like running along the bottom ?
Project Manager :	Yeah .
Industrial Designer :	You can also get the kind of the TV guide so {disfmarker}
User Interface :	It'll give you the sports .
Marketing :	Wait , is it like the Weather Channel where it's got like the ticker running on the bottom or something ?
Project Manager :	Kind of .
.	.
.	.
.	.
Project Manager :	Yep .
Marketing :	Ooh , I just got an idea for a design .
Project Manager :	gap good . Anybody have anything else they'd like to bring up in this meeting ?
Industrial Designer :	I had something , but I forgot .
Gold Summary1:	The team agreed that the buttons were big, so the older people could use them, and maintained a simple design. Industrial Designer said that the control could have a charger base with a button to find the control like a base charger of a remote phone, but Project Manager said that they could make a decision about that later. User Interface said that they included a menu button for the various things needed and for voice recognition.
Gold Summary2:	Project Manager said that speech recognition could be part of the lost-and-found function, and if they said find remote, it could beep.
Gold Summary3:	Project Manager said that everyone could agree with the clients target group, and asked if the target group was older people, but said that it would be universal for everyone. Later Marketing said that all the different age groups had different desires for speech recognition, so older people didn't care.
BART$_{large}$:	Project Manager said that the remote was only going to be a TV remote, with no TiVo, no stereo, no multi-functional, just black and white with just text on the screen. It was going to promote the company more, so they needed to include their colour and their company slogan on the remote.
T-S 1:	Project Manager said that they were going to come to some decisions, definitive, about the target group and the functions and just definite things that they needed to do and then they would close up the meeting.
T-S 2:	Project Manager said that the remote was not going to be multi-functional and that it would only be a TV remote, so no TiVo, no stereo, no DV no teletext, no TV guide, and no sports channel. Project Manager also said that no one would go to the teletext channel to check the news.
T-S 3:	Marketing said that the target group was older people. Project Manager said that it would be universal for everyone, even if it had large buttons.
T-S-L 1:	Project Manager said that they needed to come to a decision later about all the components that needed to be included, then went back to his PowerPoint to discuss the new project requirements. The new product requirements were only going to be a TV remote, with no TiVo, no stereo, and just black with just text on the screen.
T-S-L 2:	Project Manager said that they would come to a decision later about all the components that they needed to include, and then went back to his PowerPoint to discuss the new project requirements. The new product requirements were only going to be a TV remote, with no D-V-D, no TiVo, no stereo, and no multi-functional.
T-S-L 3:	Marketing said that their target group was older people. Project Manager said that even if something had large buttons, as long as they were not childishly large, non-technically challenged people would use it.
T-S-L-CL (Ours) 1:	Project Manager said that they needed to come to a decision later about all the components that needed to be included, then went back to his PowerPoint to discuss the new project requirements. New product requirements were only going to be a TV remote, not DVD, no TiVo, no stereo, black and white with just text on the screen, and promote the company more with the company color and slogan on the remote.
T-S-L-CL (Ours) 2:	Project Manager said that the remote would have a lost-and-found function to find it if it got lost. User Interface asked if it would be universal for everyone, Project Manager said no, just for older people. Project Manager added that non-technically challenged people were going to use it, so they wanted something user-friendly.
T-S-L-CL (Ours) 3:	Marketing asked if the remote control was universal for older people, Project Manager said that even if it had large buttons, non-technically challenged people would use it because they wanted something user-friendly.

Figure 8: Appendix: Sample generated summaries. Note: Topic 1 - "*ideas, function design, previous presentation*"; Topic 2 - "*lost-and-found function*"; Topic 3 - "*older people*".