ExplainMeetSum: A Dataset for Explainable Meeting Summarization Aligned with Human Intent

Hyun Kim* and **Minsoo Cho***
Superintelligence Creative Research Lab.,
Electronics and Telecommunications
Research Institute (ETRI), Republic of Korea
{h.kim, mscho}@etri.re.kr

Seung-Hoon Na†
Computer Science and Engineering,
Jeonbuk National University,
Republic of Korea
nash@jbnu.ac.kr

Abstract

To enhance the *explainability* of meeting summarization, we construct a new dataset called "*ExplainMeetSum*," an augmented version of QMSum, by newly annotating *evidence* sentences that faithfully "explain" a summary. Using ExplainMeetSum, we propose a novel *multiple extractor guided summarization*, namely *Multi-DYLE*, which extensively generalizes DYLE to enable using a supervised extractor based on human-aligned extractive oracles. We further present an explainability-aware task, named "Explainable Evidence Extraction" (E3), which aims to automatically detect all evidence sentences that support a given summary. Experimental results on the QMSum dataset show that the proposed Multi-DYLE outperforms DYLE with gains of up to 3.13 in the ROUGE-1 score. We further present the initial results on the E3 task, under the settings using separate and joint evaluation metrics.[1]

1 Introduction

Meeting summarization typically is a form of *long* document summarization, because the input is usually given as a long conversational sequence from multi-party dialogues. Among various approaches for long document summarization, the *extract-then-generate* method is one of the promising methods; it first automatically selects "salient" contents which are relevant to a specific summarization and employs them to guide the generation of a summary (Chen and Bansal, 2018; Zhang et al., 2019; Lebanoff et al., 2019; Xu and Durrett, 2019; Bajaj et al., 2021; Zhang et al., 2021; Mao et al., 2022), thereby inducing the manner of dealing with both efficiency (in processing a long input) and effectiveness (in locating accurately informative relevant contents).

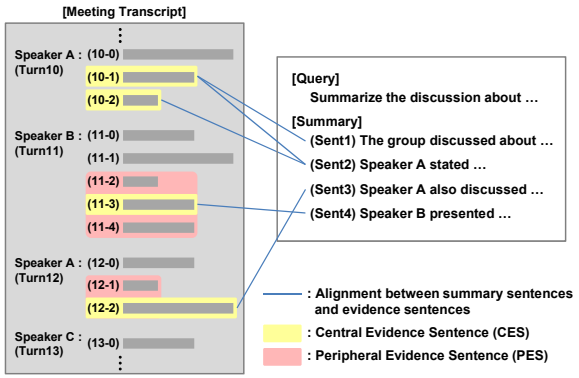

Figure 1: An illustrated example of our annotation process in ExplainMeetSum where evidence sentences are manually aligned for every single summary sentence and categorized into two types – *Central Evidence Sentence* (CES) and *Peripheral Evidence Sentence* (PES), described in Section 3.

However, the extract-then-generate method typically selects salient content in a distantly supervised or an end-to-end manner using only a final summary as a supervision signal, thereby likely being far from thoses in the chain-of-thought (or compression) required for the human summarization process. Thus, the resulting salient contents do not satisfactorily and convincingly "explain" or "support" a generated summary, and cause it to lack *explainability*.

Aiming to achieve a high degree of explainability in meeting summarization, this paper proposes a new dataset called **ExplainMeetSum**, an augmented version of QMSum, by manually and explicitly annotating *evidence* sentences that faithfully "explain" and "support" each summary sentence. Figure 1 illustrates an example of the annotation of evidence sentences. Based extensively on ExplainMeetSum, we propose Multi-DYLE, a generalized version of DYLE that enables multiple extractors, and present a novel explainability-aware benchmark task, called Explainable Evidence Extraction (**E3**), as follows.

*These authors contributed equally to this work.
†Corresponding author
[1]Our code and dataset are available at https://github.com/hkim-etri/ExplainMeetSum

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13079–13098
July 9-14, 2023 ©2023 Association for Computational Linguistics

1. **Multiple Extractors Guided Dynamic Latent Extraction for Abstractive Summarization (Multi-DYLE)** straightforwardly extends DYLE (Mao et al., 2022) by newly employing a supervised extractor trained on the evidence sentences in ExplainMeetSum in addition to the original DYLE's extractor. The underlying assumption is that, being explicitly trained using "explainable" evidence sentences, the extract-then-summarize method undertakes more likely "human-aligned" salient sentences to guide the summary generation process, potentially leading to an improvement in the quality of summaries; this effect is to some extent similar to the *chain-of-thought* prompting (Wei et al., 2022) that explicitly supervises the human's reasoning steps for the decoder in the language models.

2. **Explainable Evidence Extraction (E3)** is an explainability-aware task that aims to automatically detect all evidence sentences to explain and support a summary for meeting summarization. Thus, E3 is the task defined under the *summarize-then-explain* setting, where a generated summary is first provided and its explainable evidence sentences are extracted.

By newly employing the evidence-based supervised extractor, the experimental results on the QMSum dataset show that the proposed Multi-DYLE outperforms DYLE with an increase of 3.13 in the ROUGE-1 score. We further evaluate the baseline transformer-based models for the E3 task and present the initial experiment results under separate and joint evaluation settings that unify the meeting summarization and E3. To our best of knowledge, our work is the first to explore the explainability of meeting summarization by providing manually annotated datasets of explainable evidence sentences.

Our contributions are summarized as follows: 1) we newly introduce the *ExplainMeetSum* dataset as a valuable resource to enhance explainability in meeting summarization. 2) We propose Multi-DYLE, which enables the merging of multiple extractors in DYLE and achieves non-trivial improvements over DYLE. 3) We propose E3 using ExplainMeetSum as a new explainability-aware benchmark task, establishing the goal of extracting human-aligned explainable evidence sentences for a generated summary.

2 Related Work

2.1 Meeting Summarization

Among the various approaches for meeting summarization such as divide-and-conquer (Grail et al., 2021; Zhang et al., 2022) and hierarchical method (Zhu et al., 2020), the extract-then-summarize (or locate-and-summarize) methods have been widely adopted owing to their effective two-stage manner of handling long inputs (Chen and Bansal, 2018; Lebanoff et al., 2019; Xu and Durrett, 2019; Zhang et al., 2019; Bajaj et al., 2021; Zhang et al., 2021; Mao et al., 2022).

In particular, DYLE presented a joint training approach (Mao et al., 2022) to strengthen the interaction between the extractor and generator in a bidirectional manner by proposing a *consistency loss* that forces the extractor distribution over a set of snippets to closely match their importance degrees assigned by the generator's view.

Some studies have designed dynamic interactions between speakers during a dialogue. Qi et al. (2021) used pre-training methods based on a hierarchical encoder-decoder structure to model the semantic information between participants. Feng et al. (2020) proposed the graph modeling strategy to encode discourse relations in a conversation.

2.2 Evaluation for Extractive Summarization

Given the known limitations of using ROUGE due to its simplified n-gram matching style (Schluter, 2017), some studies have focused on evaluation in the setting of extractive summarization (Ma et al., 2021; Akter et al., 2022), pursuing automatic methods without requiring human annotation. DSMRC-S (Ma et al., 2021) transformed the summarization problem into a machine reading comprehension task, and Akter et al. (2022) proposed a *semantic-aware nCG* (normalized cumulative gain)-based evaluation metric that uses automatically generated semantic-aware ground truths.

Unlike the existing "automatic" approaches for extractive summarization, we newly present "manually" annotated ground truths and explicitly define E3 in meeting summarization, being different from the extractive summarization task.

Furthermore, evidence sentences manually extracted in our work are different from summarization content unit (SCU) (Nenkova and Passonneau, 2004; Louis and Nenkova, 2009). SCUs are obtained from multiple summaries by humans, not from an original document, whereas CES and PES

Base Dataset	QMSum		AMI	ICSI	(Total)
Query Type	General	Specific	General(long)	General(long)	
Total # of Transcripts	232	232	137	59	232
Total # of Queries	234	1576	137	59	2006
Avg. # of Sum-Sentences	5.76	3.10	18.03	22.80	5.02
Total % of Evid-Sentences (% CES / % PES)	63.04 / 36.96	75.71 / 24.29	64.90 / 35.10	67.31 / 32.69	68.98 / 31.02
Avg. # of Evid per Sum-Sent (# CES / # PES)	3.28 / 1.92	2.27 / 0.73	2.92 / 1.58	3.38 / 1.64	2.72 / 1.22

Table 1: Statistics of ExplainMeetSum. The first two rows present the total number of transcripts and queries. The third row is the average number of sentences in summaries. The fourth row is the ratio of sentences between two types of evidence where the number of CES is larger than that of PES in all sets. The last row indicates the average numbers of CES and PES per summary sentence.

are sentences (not spans) and extracted from the original meeting document, referring to only a single gold/model summary.

3 ExplainMeetSum Dataset

3.1 Annotation of Explainable Evidence Sentences

We conducted the annotation on top of the QMSum (Zhong et al., 2021), which is one of the largest datasets for meeting summarization containing "query-summary" pairs on the meeting transcripts from AMI (Carletta et al., 2006), ICSI (Janin et al., 2003), and parliamentary committee meetings. For each summary sentence, annotators were required to select aligned evidence sentences by dividing them into two types – CES and PES – according to their degrees of relevance to the query-summary pair, informally defined as follows:

Central Evidence Sentence (**CES**) is an evidence sentence with *key* information that is exactly or semantically matched with "central" parts in the summary sentence or closely related examples. An example of a CES is as follows:

(Gold Summary) The team members will work on their individual work.

(CES) Project Manager : And uh you are going to work on your individual works.

Peripheral Evidence Sentence (**PES**) is an evidence sentence that is relevant but less important than a CES, usually containing auxiliary information or examples that require a step of reasoning to match the given summary sentence. An example of a PES is as follows:

(Gold summary) The remote will have buttons for channel changing, volume settings, numerals, and power on/off.

(PES) Project Manager : but first maybe what is what are the usual function of a standard remote control?

To clearly classify evidence types, annotators were guided to choose a type of matching characteristic of a candidate evidence sentence to a summary sentence, and to determine CES for the cases of *exact*, *semantic*, and *supportive* matching types, and PES for *illustrative*, *introductory*, and *connective* matching types.

3.2 Data Collection and Statistics

Table 1 lists the statistics for the *ExplainMeetSum* dataset. The "General" and "Specific" subcolumns correspond to two types of queries in QMSum, respectively. "General(long)" subcolumn refers to the summaries of AMI and ICSI.[2]

Appendix A.2 presents samples of ExplainMeetSum with an full annotation example. Appendices A.1 and A.3 present details and quality control methods in the annotation process, respectively.

4 Multi-DYLE

Figure 2 shows the overall architecture of the Multi-DYLE model.

The key novelty of Multi-DYLE is the employment of M heterogeneous extractors with separate sets of extractive oracles,[3] M oracle losses, and a consistency loss under the generalized extractive-generator framework. This section presents the details of Multi-DYLE, including a brief description of DYLE (Mao et al., 2022).

[2]The reason to distinguishably use "General(long)" is that general summaries in AMI and ICSI tend to be longer than the general ones in QMSum.

[3]Here, our definition of extractive oracles is further generalized than that in Mao et al. (2022), while extractive oracles in Mao et al. (2022) only refer to the ROUGE-based automatically selected ones.

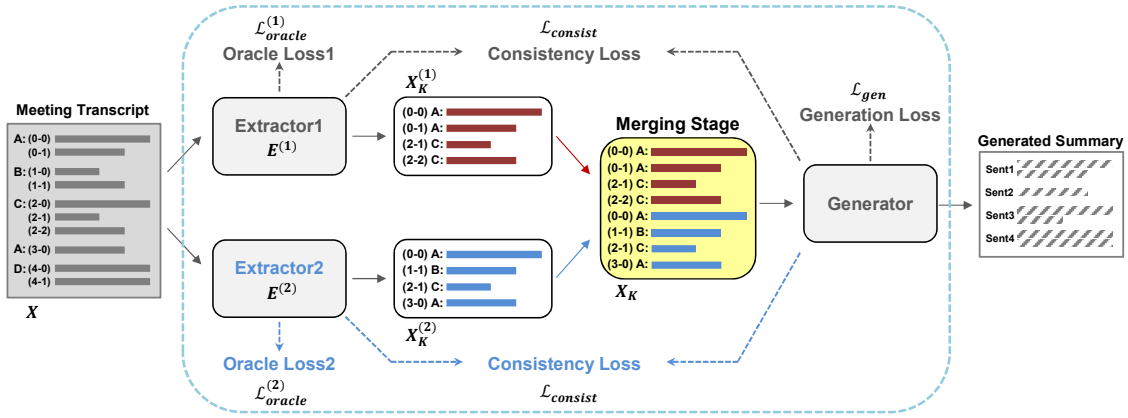

Figure 2: An overall architecture of the proposed Multi-DYLE for the case of $M = 2$: i) Multi-DYLE consists of M extractors, i.e., $\mathcal{E} = \{E^{(1)}, E^{(2)}\}$, which compute relevance scores for each sentence $x_i \in X$; ii) For the j-th extractor, we select the top K snippets with the highest relevance scores, denoted as $X_K^{(j)}$ (i.e. Eq. (1)); iii) The resulting M list of top-K sentences are then merged to obtain X_K as a final set, but allowing the "duplicated" snippets (i.e. Eq. (2)); iv) The merged set is used to guide the summary generation process at the decoding time steps (i.e., Eq. (3)); v) The dynamic weights computed by the generator over all decoding time steps are reflected back as a supervised signal to train each j-th extractor, thus leading to the consistency loss $\mathcal{L}_{consist}$ (i.e. Eq. (4)); vi) Multi-DYLE is trained with multi-task learning using the combined losses (i.e., Eq. (6))–generation loss $\mathcal{L}_{gen}$, oracle losses $\mathcal{L}_{oracle}^{(j)}$ (i.e., Eq. (5)), and consistency loss $\mathcal{L}_{consist}$.

4.1 Multiple Extractors Guided Generator

Following the notation of DYLE (Mao et al., 2022), suppose that a query q is given, and $X = (x_1, \cdots, x_L)$ is a sequence of L snippets. Unless otherwise mentioned, a snippet indicates a *single dialogue sentence* of a speaker in a meeting transcript.[4]

In contrast to DYLE that uses a "single" extractor, we have M *multiple* extractors, denoted as $\mathcal{E} = \{E^{(1)}, \cdots, E^{(M)}\}$ which computes relevance score $s_i^{(j)} = E^{(j)}(q, x_i)$ for the i-th utterance sentence x_i. For the j-th extractor, we select the top K snippets $X_K^{(j)}$ based on their relevance scores, as follows:

$$X_K^{(j)} = \text{top-K}\left(\left\{\left(x_i, E^{(j)}(q, x_i)\right)\right\}_{i=1}^{L}\right) \quad (1)$$

where top-K $(\mathcal{S})$ is the operator that chooses a list of the top K keys by sorting $\mathcal{S} = \{(a_{i1}, a_{i2})\}_{i=1}^{n}$, a set of n key-value pairs (i.e., 2-tuples) after sorting S in descending order according to their values.

The core part of Multi-DYLE is the *merging* stage, which combines the M lists of the top-K extracted sentences $\left\{X_K^{(j)}\right\}_{j=1}^{M}$ as follows:

$$X_K = X_K^{1:M} = \text{merge}\left(\left\{X_K^{(1)}, \cdots, X_K^{(M)}\right\}\right) \quad (2)$$

Our merging enables the duplicate sentences in a single list, and the same sentence is treated differently. For example, for $K = 2$ and $M = 2$, merge $(\{x_1, x_2\}, \{x_2, x_3\}) = \left\{x_1, x_2^{(1)}, x_2^{(2)}, x_3\right\}$ where $x_2^{(1)}$ and $x_2^{(2)}$ are considered differently, despite being identical.

The generator produces a summary by referring to X_K as a set of retrieved content by computing the generation probabilities $P\left(y|q, X_K\right)$, similar to an extended version of the *RAG-token* model of Lewis et al. (2020), as follows:

$$P\left(y|q, X_K\right) = \prod_{t=1}^{T} \sum_{x \in X_K} P\left(x|q, X_K, y_{1:t-1}\right) P\left(y_t|q, x, y_{1:t-1}\right) \quad (3)$$

where $y_{1:t-1}$ is the previously generated sequence at the t-th decoding time step, $P\left(x|q, X_K, y_{1:t-1}\right)$ is the *dynamic* weight of the snippet x, and $P\left(y_t|q, x, y_{1:t-1}\right)$ is the generation probability when x is used as the additional encoded context.

Similar to DYLE, Multi-DYLE uses the *average* of the dynamic weights of a sentence $x \in X_K$ across T time steps as a *supervised signal* to train

[4]Our "sentence"-level setting is different from DYLE, which uses a "turn"-level snippet to refer to an utterance that usually consists of multiple sentences, as described in Appendix B.

M extractors, thereby introducing *consistency loss*, as follows:

$$\mathcal{L}_{consist} = \mathsf{KL}\Bigg[\frac{1}{T}\sum_{t=1}^{T} P(\cdot|q, X_K, y_{1:t-1}) \Big\| \mathsf{softmax}\left(E(q, x_i), x_i \in X_K\right)\Bigg] \quad (4)$$

where $E(q, x_i) = E^{(j)}(q, x_i)$ when x_i belongs to the top sentences selected by the j-th extractor, i.e., $x_i \in X_K^{(j)}$.

4.2 Multiple Extractive Oracles

To provide basic supervised signals for multiple extractors, we employ M separate *extractive oracles*, $\left\{X_o^j\right\}_{j=1}^{M}$, thus introducing M *oracle losses*, defined as follows:

$$\mathcal{L}_{oracle}^{(j)} = -\frac{1}{|X_o^{(j)}|}\sum_{x \in X_o^{(j)}} \log \frac{e^{E^{(j)}(q,x)}}{\sum_{i=1}^{L} e^{E^{(j)}(q,x_i)}} \quad (5)$$

In our setting, we deploy two different sets of extractive oracles for $X_o^{(j)}$: ROUGE-based extractive oracles, as in DYLE, and our CES-based extractive oracles, which we clearly specify in Section 6.1.1.

4.3 Generalized Training Objective

The final training objective is based on the M oracle losses and the consistency loss as follows:

$$\mathcal{L} = \lambda_g \mathcal{L}_{gen} + \lambda_o \sum_{j=1}^{M} \mathcal{L}_{oracle}^{(j)} + \lambda_c \mathcal{L}_{consist} \quad (6)$$

where $\mathcal{L}_{gen}$ is the *generation* loss using NLL defined in DYLE (Mao et al., 2022), and λ_g, λ_o and λ_c are hyperparameters, which are fixed to 1 in this study.

Multi-DYLE degenerates to DYLE when $M = 1$ using $X_o^{(1)}$ as a set of ROUGE-based extractive oracles.

5 Explainable Evidence Extraction (E3)

In this section, we introduce the details of E3, which identifies all CESs and PESs for a given summary, and baseline E3 models.

5.1 Task Definition

Different from the summarization task in Section 4, we now have a summary S, given as a sequence of N *summary sentences* $S = (s_1, \cdots, s_N)$ where S is either a gold summary or automatically generated one.

Given the meeting transcript $X = (x_1, \cdots, x_L)$, let $Y_k \subseteq X$ be a ground-truth set of CESs and PESs for the k-th summary sentence $s_k \in S$, obtained in ExplainMeetSum. E3 is thus defined as the task of automatically identifying Y_k for a given $s_k \in S$.

5.2 Model

As our baseline E3 model, often referred to as the *evidence extractor (EE)*, we employ the extractor module in the DYLE (Mao et al., 2022) model, but using a given summary sentence as an additional input for the encoder. Formally, the EE's input is a concatenated sequence of the k-th summary sentence s_k, query q, and meeting transcript X, presented as (s_k, q, X). EE then produces relevance scores for the i-th sentence $x_i \in X$.[5]

Because the meeting transcript is often too long to be contained within the maximum length limit, we split the transcript into a list of "chunks" with the fixed size of tokens, and separately encode all the chunks. The relevance score of the i-th sentence x_i is obtained from the chunk-level representation which x_i belongs to.

For training, the cross-entropy loss is adopted to maximize the classification probability of gold evidence sentences in the CES and PES. For the inference time, we further apply a filtering step to the classification probabilities, using *threshold*-based and *top-K* selection methods, as discussed in Section 6.2.

6 Experiments

In our experiment, we first compare the summarization performance of Multi-DYLE, introduced in Section 4, with that of DYLE and its simple variants to check whether the use of multiple extractors lead to performance improvement. We further present the performance of our baseline EE described in Section 5 under the settings of separate and joint tasks in Sections 6.2 and 6.3, respectively. An illustration and examples of joint tasks are describe in Appendix C, and the implementation details for the Multi-DYLE and EE models are presented in Appendix D.

[5] More precisely saying, Mao et al. (2022) appended the special token </s> between x_{i-1} and x_i and computed the output score from the token's output representation.

6.1 Meeting Summarization

6.1.1 Main Results

Table 2 presents the comparison results of Multi-DYLE (i.e. using ExplainMeetSum) and DYLE for QMSum.

As aforementioned in Section 4, Multi-DYLE uses sentence-level snippets whereas the original version of DYLE uses turn-level snippets.

To clarify the different setups for using extractive oracles, with the abuse of notation, X_o^{ROG}, X_o^{CES}, and X_o^{PES} refer to the sets of ROUGE-based, CES-based, and PES-based extractive oracles (in ExplainMeetSum), respectively. The various types of Multi-DYLE are defined as follows:

- **Multi-DYLE**(X_o^{α}): the run using a single extractor ($M = 1$) based on $X_o^{(1)} = X_o^{\alpha}$
- **Multi-DYLE**($X_o^{\alpha}, X_o^{\beta}$): the run using dual extractors ($M = 2$) based on $X_o^{(1)} = X_o^{\alpha}$ and $X_o^{(2)} = X_o^{\beta}$.

Some variants of DYLE using X_o^{ROG} are denoted as follows:

- **DYLE**(X_o^{ROG}): the variant of DYLE using the fine-tuned DYLE model at turn-level settings[6] but applying it to sentence-level snippets at inference time.
- **Multi-DYLE**(X_o^{ROG}): the variant of DYLE in which both fine-tuning and testing are conducted under our setting of *sentence-level* snippets, unlike DYLE(X_o^{ROG}).[7]

Interestingly, by performing inference only at sentence-level utterances without any fine-tuning, DYLE(X_o^{ROG}) achieves a ROUGE-1 of 35.41, with an increase of approximately 1 in ROUGE-1 over the original turn-level DYLE. Multi-DYLE(X_o^{ROG}) further increases the performance by fully fine-tuning DYLE in the sentence-level setting. The results consistently show that sentence-level snippets are more effective than turn-level ones.

Using the CES-based extractive oracles X_o^{CES}, it is noticeable that Multi-DYLE(X_o^{CES}) further improves the performance of Multi-DYLE(X_o^{ROG}), resulting in an increase of about 0.7 in ROUGE-1.

[6]This DYLE model is made publicly available by the authors of DYLE.

[7]The turn-level and sentence-level oracle examples can be found in Appendix B.

Model	ROUGE R-1	R-2	R-L
<Baselines>			
BART-LS (Xiong et al., 2022)	37.9	12.1	33.1
SecEec-W (Vig et al., 2022)	37.80	13.43	33.38
DYLE (Mao et al., 2022)	34.42	9.71	30.10
<Ours - Sentence level>			
DYLE (X_o^{ROG}) with turn-level finetuning	35.41	10.74	31.00
Multi-DYLE (X_o^{ROG}) ⓐ	35.93	11.24	31.26
Multi-DYLE (X_o^{CES}) ⓑ	36.63	11.81	31.82
Multi-DYLE (X_o^{ROG}, X_o^{CES}) ⓒ	37.55	12.43	32.76

Table 2: Meeting summarization results on test sets of QMSum, comparing Multi-DYLE and DYLE with other previous works, under ROUGE scores as evaluation metrics.

By merging the two types of extractors, Multi-DYLE(X_o^{ROG}, X_o^{CES}) leads to non-trivial improvements over runs with a single extractor (i.e., Multi-DYLE(X_o^{ROG}) or Multi-DYLE(X_o^{CES})), finally achieving 37.55 of ROUGE-1.

Overall, the results confirm that the use of human annotated evidence sentences improves performance under the same framework, even without changing the model's architecture.

6.1.2 Ablation Study

Table 3 presents the ablation study of Multi-DYLE with different setups of extractive oracles by varying the number of extracted sentences K. In addition to the runs in Table 2, we consider the union of two sets of extractive oracles – $X_o^{ROG} \cup X_o^{CES}$ and $X_o^{CES} \cup X_o^{PES}$.

In Table 3, the last two columns named "ROG-Oracle" and "CES-Oracle" refer to the *extraction performances* of $\left\{X_K^{(j)}\right\}_{j=1}^{M}$, which are selected by M extractors (i.e., using Eq. (1) and their merged results X_K (that is using Eq. (2)) under the precision, recall, and F1 metrics when using X_o^{ROG} and X_o^{CES} as ground-truth sets, respectively. Here, the subcolumns named "Ext1," "Ext2," and "Merged" indicate the extraction results of $X_K^{(1)}$, $X_K^{(2)}$, and X_K, respectively.

It is clearly shown that the ROUGE scores tend to be proportional to the F1 score of the extraction when either X_o^{ROG} or X_o^{CES} is the ground-truth set. Particularly, the ROUGE scores are slightly more proportional to F1 when X_o^{ROG} is a ground-truth set, compared to the case that uses X_o^{CES} as the gold standard. When $M = 1$, in the "ROG-Oracle" subcolumn, an interesting result is that Multi-DYLE($X_o^{ROG} \cup X_o^{CES}$) achieves the best F1 result and ROUGE score, meaning

M	Multi-DYLE	# Top-K Ext1	Ext2	Merged	ROUGE ↑ (R-1/R-2/R-L)	ROG-Oracle (P/R/F1) Ext1	Ext2	Merged	CES-Oracle (P/R/F1) Ext1	Ext2	Merged
1	Multi-DYLE ($\mathsf{X}_o^{\mathsf{ROG}}$) ⓐ	30	-	30	35.93/ 11.24/ 31.26	8.22/ 40.89/ 12.81	-	8.22/ 40.89/ 12.81	8.00/ 37.96/ 12.55	-	8.00/ 37.96/ 12.55
	Multi-DYLE ($\mathsf{X}_o^{\mathsf{CES}} \cup \mathsf{X}_o^{\mathsf{PES}}$)	30	-	30	36.49/ 11.56/ 31.67	8.35/ 44.95/ 13.43	-	8.35/ 44.95/ 13.43	11.55/ 51.46/ 17.82	-	11.55/ 51.46/ 17.82
	Multi-DYLE ($\mathsf{X}_o^{\mathsf{CES}}$) ⓑ	30	-	30	36.63/ 11.81/ 31.82	8.16/ 45.10/ 13.21	-	8.16/ 45.10/ 13.21	11.77/ 52.28/ 18.13	-	11.77/ 52.28/ 18.13
	Multi-DYLE ($\mathsf{X}_o^{\mathsf{ROG}} \cup \mathsf{X}_o^{\mathsf{CES}}$)	30	-	30	36.93/ 12.18/ 32.49	9.25/ 47.56/ 14.65	-	9.25/ 47.56/ 14.65	10.89/ 49.33/ 16.87	-	10.89/ 49.33/ 16.87
2	Multi-DYLE ($\mathsf{X}_o^{\mathsf{ROG}}$, $\mathsf{X}_o^{\mathsf{CES}} \cup \mathsf{X}_o^{\mathsf{PES}}$)	15	15	30	37.10/ 12.19/ 32.56	11.55/ 31.13/ 15.46	10.58/ 30.86/ 14.81	9.84/ 42.28/ 15.11	11.53/ 28.54/ 15.33	16.16/ 38.10/ 21.10	12.03/ 44.23/ 17.98
	Multi-DYLE ($\mathsf{X}_o^{\mathsf{ROG}}$, $\mathsf{X}_o^{\mathsf{CES}}$) ⓒ	15	15	30	37.55/ 12.43/ 32.76	12.55/ 31.92/ 16.63	10.49/ 30.33/ 14.55	10.23/ 41.82/ **15.53**	12.05/ 29.84/ 16.01	16.92/ 39.43/ 21.93	12.66/ 45.97/ **18.79**

Table 3: Meeting summarization results of Multi-DYLE (X_o^{α}) when $M = 1$ and Multi-DYLE $\left(\mathsf{X}_o^{\alpha}, \mathsf{X}_o^{\beta}\right)$ when $M = 2$ on test sets of QMSum, varying different settings of extractive oracles X_o^{α} and X_o^{β} under ROUGE scores as summarization metric (4th column, named ROUGE), and P/R/F1 scores as the extraction performance when $\mathsf{X}_o^{\mathsf{ROG}}$ (5th column, named ROG-Oracle) and $\mathsf{X}_o^{\mathsf{CES}}$ (6th column, named CES-Oracle) are used as gold sets.

that when X_o^{CES} is used for an additional training set of an extractor, it has a positive impact on extracting X_o^{ROG}. Although Multi-DYLE(X_o^{CES}) shows weak performance in correctly extracting X_o^{ROG}, it shows better ROUGE scores than Multi-DYLE(X_o^{ROG}).

When $M = 2$, Multi-DYLE(X_o^{ROG}, X_o^{CES}) slightly improves the performance of Multi-DYLE(X_o^{ROG}, $X_o^{CES} \cup X_o^{PES}$), indirectly indicating that the additional use of X_o^{PES} for training an extractor does not lead to further improvement in summarization.

Overall, the results confirm that a strong correlation exists between ROUGE and F1 scores, enabling us to reasonably predict whether the model improves, based on the F1 scores of the extractors. The extractor trained only on X_o^{ROG} does not exhibit the best performance in extracting X_o^{ROG}, whereas the additional use of X_o^{CES} is complementary in identifying X_o^{ROG}.

6.2 Explainable Evidence Extraction

Table 4 compares the results of our baseline EE model for E3, described in Section 5.2. Here, we use the *gold* setting in which a gold summary is assumed to be provided for EE.

When extracting evidence sentences for each summary sentence, we have three types of filtering methods based on the classification probabilities for all candidate sentences in the meeting transcript $X = (x_i)_{i=1}^{L}$:

- *Threshold-based* method (i.e., thr-θ): selects a sentence as an evidence sentence when its classification probability is larger than the threshold θ
- *Top-R* method (i.e., top-R): selects R sentences with the highest classification probabilities
- *Hybrid* method (i.e., thr-θ&top-R): first applies thr-θ and then conditionally performs top-R when no sentence with thr-θ is selected. thr-1.0&top-R is equivalent to top-R.

As shown in Table 4, in terms of the sentence-level E3 metric (in the upper part), the threshold-based methods (i.e. thr-θ) show higher F1 scores than the top-R methods (i.e. top-R).[8] Despite its superiority, thr-θ often suffers from its low recall; in our preliminary analysis, we observed the cases where no sentence was selected, given a threshold. Given that top-R is relatively strong in terms of the recall metric, the hybrid method (i.e., thr-θ&top-R) further increases performances, leading to achieve the best F1 score of 52.91. Similar results are observed in the summary-level metric in the lower part of Table 4; the hybrid method shows the best performance, outperforming both individual methods of top-R and thr-θ, although the results are not fully presented.

[8] In Table 4, as mentioned in its caption, the "sentence"-level P/R/F1 scores for E3 are computed based on the the gold sets are defined per summary "sentence," resulting in the *two-stage macro-averaged* score, i.e., the sentence-level scores are first averaged per query and then further marco-averaged across queries. The "summary"-level P/R/F1 scores for E3 use the gold sets defined per "summary," resulting in the *single-stage macro-averaged* score, i.e. the summary-level scores are macro-averaged across queries.

Sentence-level Macro-averaged E3 Evaluation						
EE Model	# Evidence Extraction			Sentence-level		
	Mean	STD	diff	P	R	**F1**
thr-0.5	3.35	2.26	0.42	53.98	54.43	49.31
thr-0.9	2.99	2.00	0.78	53.91	52.42	48.55
top-5	5.00	0.00	1.23	37.76	62.90	43.16
top-10	10.00	0.00	6.23	23.42	73.04	33.00
thr-0.5&top-5	3.99	2.13	0.22	56.94	59.28	52.75
thr-0.9&top-5	3.76	1.86	0.01	57.76	58.45	**52.91**
ExplainMeetSum	3.77	2.04				

Summary-level Macro-averaged E3 Evaluation						
EE Model	# Evidence Extraction			Summary-level		
	Mean	STD	diff	P	R	**F1**
thr-0.9	13.01	15.67	4.85	59.83	48.62	49.94
thr-0.9&top-5	16.60	19.12	1.26	53.24	55.35	**51.63**
ExplainMeetSum	17.86	25.72				

Table 4: E3 results of the baseline EE model on test sets of the gold evidences in ExplainMeetSum, varying the filtering options among the threshold-based selection (thr-θ), the top-R selection (top-R), and its hybrid method (thr-θ&top-R). 1) Upper: *Sentence*-level P/R/F1 scores where ground-truth CESs and PESs are defined per gold summary "sentence." 2) Lower: *Summary*-level P/R/F1 scores, where ground-truth CESs and PESs are defined per gold "summary."

In Table 4, the subcolumns in "# Evidence Extraction" named 'Mean' and 'STD' indicate the mean and standard deviation of the number of the extracted evidence sentences, respectively, and 'diff' refers to the absolute difference between the mean numbers of extracted evidence sentences and gold ones. It is clearly seen that 'diff' strongly correlates with the F1 scores of the E3 extractor.

Overall, the results show that the hybrid method produces the best F1 scores and these performances are correlated with how closely the number of extracted sentences is distributed to that of gold sentences.

6.3 Joint Evaluation of Summarization and E3

In this section, we evaluate a pipelined model consisting of Multi-DYLE and the baseline EE model. To jointly evaluate a generated summary and its aligned evidence sentences in a single metric, we adopt ROUGE scores by merely viewing the addressed task as a unified sequence "generation" task.

To be more specific, we first obtain a unified sequence from a summary and its aligned evidence sentences. With abuse of notation, suppose that $S = (s_1, \cdots, s_N)$ is a given summary and $\mathcal{X} = \left(X_e^{(i)}\right)_{i=1}^{N}$ is a list of N evidence sentence collections where $X_e^{(i)} = \left(x_1^{(i)}, \cdots, x_{m_i}^{(i)}\right)$ is a sequence of m_i evidence sentences aligned to explain the i-th summary sentence $s_i \in S$. The conversion process $\psi(S, \mathcal{X})$ is defined as follows:

$$\psi(S, \mathcal{X}) = \bigoplus_{i=1}^{N} \left(s_i \oplus x_1^{(i)} \oplus \cdots \oplus x_{m_i}^{(i)} \oplus \texttt{"\textbackslash n"} \right) \quad (7)$$

where $\oplus$ is the concatenation operator and "\n" indicates a newline character, making $\psi(S, \mathcal{X})$ comprise of N sentences.

Under this conversion process ψ, we compute ROUGE scores by matching an output sequence resulting from the application of Multi-DYLE and the baseline EE model with its corresponding ground-truth sequence. To distinguish ROUGE scores in meeting summarization, we use *Joint ROUGE* scores to indicate the unified ROUGE scores evaluated in the joint setting.

[Sum. & E3 Models]		[Sum. Task]	[Joint Sum. & E3 Tasks]	
		ROUGE	Summary -level	Joint ROUGE
Multi-DYLE	**EE**	**R-1**/ R-2/ R-L	P/ R/ **F1**	R-1/ R-2/ **R-L**
(X_o^{ROG}) ⓐ	thr-0.9 &top-5	35.93/ 11.24/ 31.26	17.09/ 20.08/ 16.45	45.51/ 21.25/ 37.74
(X_o^{CES}) ⓑ		36.63/ 11.81/ 31.82	17.87/ 23.73/ 18.46	46.63/ 22.69/ 39.04
(X_o^{ROG}, X_o^{CES}) ⓒ		37.55/ 12.43/ 32.76	20.31/ 26.01/ **20.48**	47.45/ 24.06/ **40.04**

Table 5: Results for meeting summarization and E3 on the QMSum dataset, under ROUGE scores as the summarization metric, the extraction scores of E3 and the Joint ROUGE score as the joint evaluation metric.

Table 5 presents comparison results of the pipelined system of Multi-DYLE and the baseline EE model with the hybrid selection method thr-0.9&top-5, varying sets of extractive oracles, evaluated by Joint ROUGE, as well as the evaluation metrics for meeting summarization and E3. It should be noted that the evaluation of E3 in Table 5 is the (indirectly) joint setting based on automatically generated summaries by Multi-DYLE, unlike the gold setting in Table 4.

The results show that Multi-DYLE(X_o^{ROG}, X_o^{CES}) with dual extractors again achieves the best performance under the joint settings involving E3, exhibiting increases of approximately 2 in the Joint ROUGE scores and of approximately 4 in the F1 score of E3, compared to those of

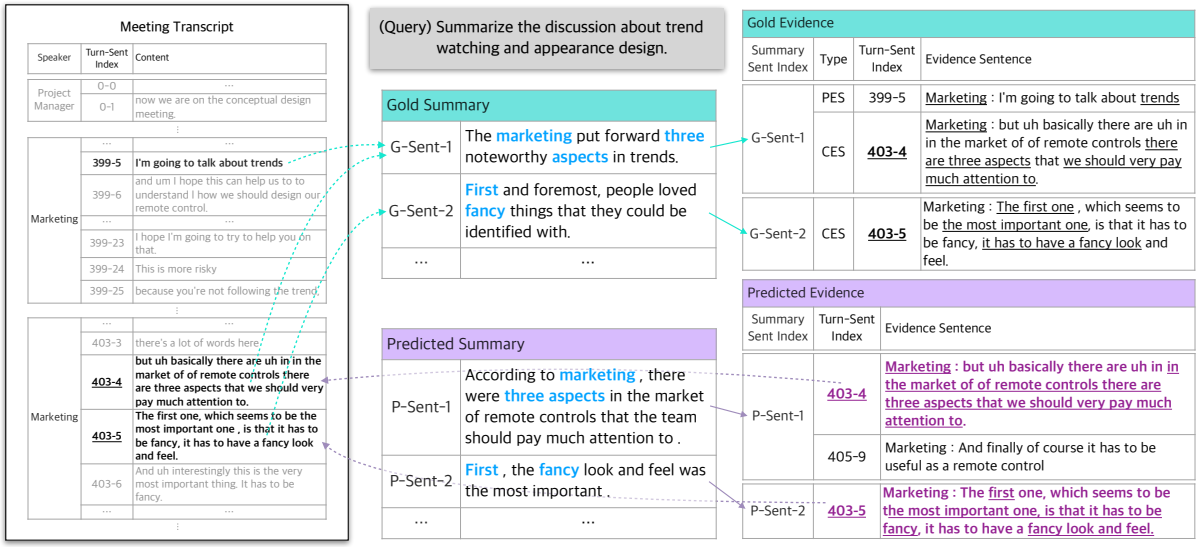

Figure 3: Example of ExplainMeetSum of comparing the summary and evidence sentences between the predicted results and gold ones for the query in QMSum (i.e., the transcript name: IS1006c, the query id: specific-6); G-Sent-n/P-Sent-n refer to n-th gold/predicted summary sentence. More examples of the summary sentences with the extracted evidence sentences for the same query are presented in Tables 8 and 13.

Multi-DYLE(X_o^{ROG}). The performance gain obtained through both Multi DYLE(X_o^{ROG}, X_o^{CES}) and the EE model tends to be further enlarged compared to the gain of ROUGE only by Multi-DYLE(X_o^{ROG}, X_o^{CES}) for meeting summarization.

Overall, Multi-DYLE(X_o^{ROG}, X_o^{CES}) shows the best performance on all evaluation metrics in meeting summarization and E3, achieving noticeable improvements particularly at joint settings involving E3.

6.4 Case Study

As an illustrated example for the query "*Summarize the discussion about trend watching and appearance design*" in QMSum, Figure 3 presents some of the predicted summary sentences using Multi-DYLE(X_o^{ROG}, X_o^{CES}) (i.e., P-Sent-n), and their evidence sentences extracted using the hybrid EE model with thr-0.9&top-5, in comparison with the gold summary sentences (i.e., G-Sent-n) and their human-aligned CESs. As shown in Figure 3, the generated summary sentences, P-Sent-1∼2 are semantically matched well with the gold summary ones, (i.e., G-Sent-1∼2), by including the common keywords such as "marketing," "three aspects," "first," and "fancy." Importantly, all CESs (i.e., 403-4 and 403-5) for the gold summary are correctly extracted in the predicted evidence sentences for P-Sent-1∼2, confirming that P-Sent-1∼2 is a high quality summary which is supported by the human-aligned CESs.

7 Conclusion

In this paper, we presented a novel ExplainMeetSum, as an explainability-enhanced QMSum by providing complete manual annotation of two types of evidence sentences, CES and PES, as explanations to faithfully support or explain each sentence in gold summaries. Equipped with ExplainMeetSum, we proposed Multi-DYLE as a generalized DYLE to enable the addition of an explainable extractor based on CES-based extractive oracles. We further defined a novel task, E3, which aims to extract explainable evidence sentences when a summary sentence is given. The experimental results obtained on QMSum using ExplainMeetSum showed that the proposed Multi-DYLE based on an additional extractor towards a human-aligned explanation outperformed DYLE and led to improvements in the joint evaluation settings involving E3.

In future work, we would like to invent a joint learning framework for meeting summarization and E3, extensively employing human-supervised explainable signals from ExplainMeetSum, towards better explainable meeting summarization. Furthermore, developing a novel joint evaluation metric for meeting summarization and E3 to overcome the limitations of the ROUGE-driven scores would be worthwhile.

Limitations

This paper presents ExplainMeetSum to enhance the explainability of meeting summarization and provides Multi-DYLE as a generalized version of DYLE by employing multiple extractors. One limitation of our work is the restricted exploration of using ExplainMeetSum for meeting summarization. Although we propose the use of multiple extractors, it can go beyond DYLE's extractive-generator framework, thereby extending and generalizing other extract-then-generate methods, such as Dou et al. (2021). In addition, we currently use single and dual extractors (i.e., $M = 1$ or $M = 2$) for Multi-DYLE. However, other advanced settings using more extractors ($M > 2$) were not examined in this experiment.

Another limitation is the current joint evaluation metrics, such as precision, recall, F1 scores, and Joint ROUGE scores, adopted as initial trials of evaluating of the "summarize-then-explain" joint setting. In particular, Joint ROUGE scores inherit the limitations of the original ROUGE scores. An important remaining issue is to design a more stable and agreeable joint evaluation metric that can be used as a standard evaluation metric for the joint setup of the summarize-then-explain task.

Furthermore, our current applications using ExplainMeetSum are limited to meeting summarization and E3. However, ExplainMeetSum can be used as a suitable benchmark dataset to compare various interpretable and explainable models on summarization results. Given the emerging importance of interpretable and explainable models, arguably valuable work is to extensively examine the usefulness of ExplainMeetSum in more interpretable-related tasks by exploring and evaluating interpretable models, such as Ribeiro et al. (2016); Lundberg and Lee (2017); Sundararajan et al. (2017); Sanyal and Ren (2021); Saha et al. (2022).

Some parts of the annotations in ExplainMeetSum were not fully utilized in this work. We also annotated the evidence sentences for General(long) types of queries in AMI and ICSI, however, our work on meeting summarization used only QMSum as a benchmark dataset. Thus, it would be valuable to obtain additional results using ExplainMeetSum for meeting summarization in AMI and ICSI to examine whether the use of ExplainMeetSum leads to improvements in other types of datasets.

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2022-0-00989,Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling) and (2019-0-00004,Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners)

We would like to thank all anonymous reviewers for their valuable comments and suggestions.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE? In Findings of the Association for Computational Linguistics: ACL 2022, pages 1547–1560, Dublin, Ireland. Association for Computational Linguistics.

Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das, and Andrew McCallum. 2021. Long document summarization in a low resource setting using pretrained language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 71–80, Online. Association for Computational Linguistics.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In Machine Learning for Multimodal Interaction, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

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

Technologies, pages 4830–4842, Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization. volume abs/2012.03502.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing BERT-based transformer architectures for long document summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1792–1810, Online. Association for Computational Linguistics.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., volume 1, pages I–I.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Advances in Neural Information Processing Systems, volume 33, pages 9459–9474. Curran Associates, Inc.

Annie Louis and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 306–314, Singapore. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Bing Ma, Cao Liu, Jingyu Wang, Shujie Hu, Fan Yang, Xunliang Cai, Guanglu Wan, Jiansong Chen, and Jianxin Liao. 2021. Distant supervision based machine reading comprehension for extractive summarization in customer service. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, page 1895–1899, New York, NY, USA. Association for Computing Machinery.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022. DYLE: Dynamic latent extraction for abstractive long-input summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1687–1698, Dublin, Ireland. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

MengNan Qi, Hao Liu, YuZhuo Fu, and Ting Liu. 2021. Improving abstractive dialogue summarization with hierarchical pretraining and topic segment. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1121–1130, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal. 2022. Summarization programs: Interpretable abstractive summarization with neural modular trees.

Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10285–10299, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 3319–3328. JMLR.org.

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Wenhan Xiong, Anchit Gupta, Shubham Toshniwal, Yashar Mehdad, and Wen-tau Yih. 2022. Adapting pretrained text-to-text models for long text sequences.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3292–3303, Hong Kong, China. Association for Computational Linguistics.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 789–797, Hong Kong, China. Association for Computational Linguistics.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4426–4433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 194–203, Online. Association for Computational Linguistics.

A Dataset Construction

A.1 Annotation Process

ExplainMeetSum was built on top of the QMSum dataset. Here, we choose QMSum because its dataset is one of the most widely used datasets and its further annotated dataset, ExplainMeetSum, is also invaluable, which is likely to be of wide research interest. For annotation, we recruited four annotators and a coordinator based on their proficiency in English and prior experience with English dataset annotations. To create ExplainMeetSum, they were required to select and annotate evidence sentences from transcripts that are considered as aligned to "explain" well each gold sentence summary in the approximately 2,000 queries distributed among 232 transcripts; once selecting an evidence sentence, an annotator labeled its main type, CES and PES, as well as matching characteristics (i.e., subtypes of CES and PES).

Despite QMSum being relatively limited in its size, the annotation work is time-consuming because the annotation tasks are non-trivial and involve extensive revisions, based on multiple feedbacks and comments provided by the coordinator. Considering the large amount of annotation work, we set aside more than six months for this task based on the two-level feedback pipeline on which the annotators interacted with both a coordinator and an expert. The coordinator thoroughly checked the annotation process and interacted with the annotators, while the expert acted as a meta-reviewer who periodically inspected random samples in-depth, provided feedback, and updated the guidelines. The total cost of the annotator's work was $40,000 which equates to an estimation of $20 per query.

A.2 Examples in ExplainMeetSum

Table 7 shows examples in ExplainMeetSum collected across various queries in QMSum. Being categorized with main types and matching characteristics (i.e., subtypes), CESs and PESs aligned for each of the gold summary sentences indicated by '[G-Sent]' are presented in the column 'Evidence Sentence.'

Matching characteristics are classified into six labels: CES has *exact*, *semantic*, and *supportive* subtypes, whereas PES has *illustrative*, *introductory*, and *connective* subtypes, defined as follows:

- **CES/exact**: A subtype of CES whose keywords are "exactly" matched with the contents in the target sentence of the summary.

- **CES/semantic**: A subtype of CES whose keywords "semantically" match the contents in the target sentence of the summary.

- **CES/supportive**: A subtype of CES that includes information or examples that directly support or entail abstractive expressions in the target sentence of the summary.

- **PES/illustrative**: A subtype of PES that includes relevant information or examples, such as enumerating further detailed information in addition to CES, on which a reasoning step is required to match the target sentence of the summary.

- **PES/introductory**: A subtype of PES, which includes relevant information that initiates the subject of the extracted evidences; it usually appears before CES.

- **PES/connective**: A subtype of PES that includes relevant information used to build the connectivity between the extracted evidences; it is usually based on conjunctions that connect two subsequent CESs.

As a full annotation example, Table 8 presents CESs and PESs for the query "*Summarize the discussion about trend watching and appearance design*" in QMSum (i.e., the transcript name: IS1006c, the query id: specific-6), where the 'Type/subtype' column shows the main type of evidence sentences (i.e., CES or PES) and their matching characteristic, the 'Evidence Sentences' column presents the aligned CESs or PESs in the meeting transcript, and the 'Turn-sent Index' column shows the *turn-based sentence identifier* that is defined as a pair of turn and sentence-level indexes.[9]

A.3 Quality Control Methods

To access the quality of the annotation, we established a two-level feedback-driven annotation process that was supervised and monitored by a coordinator and an expert (as a meta-reviewer) as follows:

- *Two-level coordinator-expert feedback process*: 1) a coordinator periodically checks the annotation results such that they are continuously revised to sufficiently fulfill the high-level of quality, and frequently communicates by an expert. 2) an expert regularly inspects random samples of the annotation, provides feedback to the coordinator, and updates the guideline.

We further applied a series of test suites to check the annotation quality semi-automatically, as follows:

- *Comparing labeling statistics across annotators*: We periodically compare labelling statistics from annotators, and reexamine annotation results when some statistics are significantly different from the others.

- *Computing neural similarities between aligned evidence sentences and summary ones*: We compare similarities using the SBERT model between candidate sentences and a summary sentence, under the assumption that the exact, semantic, supportive CES, PES, and others have the highest degree of similarity in that order. For the SBERT model, we used the sentence transformer model from Hugging Face's 'all-MiniLM-L6-v2' model to compute similarities between candidate sentences and a summary sentence.

  Table 9 presents examples of how initial tagging errors are revised correctly via the test-suite based on neural similarities between evidence and summary sentences. As in the table, given evidence sentences initially labeled by annotators, the coordinator is further provided with the SBERT-based similarities between evidence sentences and the gold summary sentences. When their similarities are abnormally large or small compared to the average similarity values of their subtypes, the coordinator re-examines the abnormal cases and revises them correctly. In the first row, for example, an annotator initially labelled the type "CES/supportive" for the given evidence sentence. However, its

[9] Supposing that x-y is the value in 'Turn-sent index,' x and y refers to the *turn-level* and *sentence-level* indices, respectively. For example, 399-5 refers to the index of 5-th sentence in the 399-th turn in the meeting transcript.

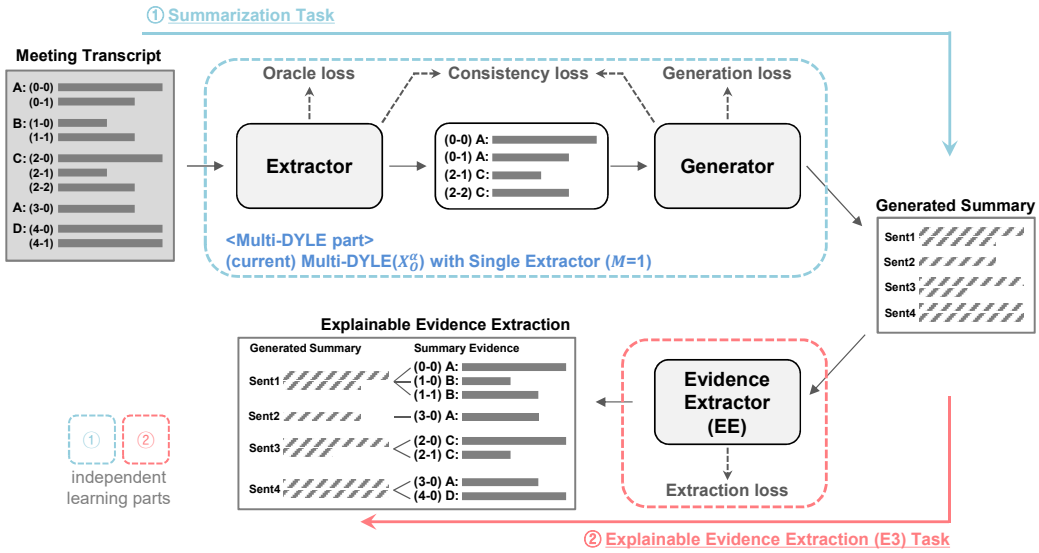

Figure 4: Overall framework of the proposed architecture for meeting summarization and E3 tasks. Multi-DYLE in Section 4 and the EE model in Section 5 are deployed to perform meeting summarization and E3 tasks, respectively, where they are trained independently.

SBERT-based similarity with the gold one is 72.20, which is "abnormally" high considering the average of CES/supportive-typed cases. Once this abnormal neural similarity was detected and alarmed, a coordinator carefully looked into the problematic evidence sentence, identified that its subtype was wrongly labeled, and finally, correctly revised the label to "CES/semantic." Similar revisions were made in other two rows.

- *Labeling matching characteristics*: In our work, CES and PES are the main types of evidence sentences only required for meeting summarization and E3 tasks, whereas matching characteristics were not necessary for these tasks. Importantly, our intention on labelling matching characteristics (i.e., subtypes of CES and PES) is to provide an additional quality control suite, thus making annotator carefully examine sentences in more depth to reduce errors in annotating CES and PES.

B Extractive Oracles: Turn-level and Sentence-level

In Table 2, we evaluated DYLE based on two types of extractive oracles — turn-level and sentence level ones – and showed that DYLE equipped with sentence-level extractive oracles exhibited improvements over the case using turn-level oracles. To clearly demonstrate the difference between these types of extractive oracles, Tables 10 and 11 show the turn-level and sentence-level ROUGE-based extractive oracles obtained for the same query in Table 8, where bold-faced index refers to human-annotated aligned CESs. It is clearly seen that the resulting sentences are different in in two types of oracles; while some turns, including simple ones that consist of a single sentence, appear in both turn-level and sentence-level oracles (i.e., 392-nd, 405-th, 427-th, and 438-th turns), most turns are not shared across them. In particular, even if a turn is shared between two oracles, a small number of sentences tend to commonly appear, as in the 405-th turn and its sentences.

C Summarization and E3 Tasks

Figure 4 presents the overall framework of the proposed architecture that performs meeting summarization and E3 tasks. In the upper part, the Multi-DYLE in Section 4 is deployed to perform the summarization task, where the Multi-DYLE of $M = 1$ can be replaced with other variants of Multi-DYLE(X_o^α, X_o^β). In the lower part, the EE model in Sections 5 and 6.2 is employed to address the E3 task. Although Multi-DYLE includes its evidence extractor, we used a separate EE model to address the E3 task. During inference, given

a test query and meeting script, Multi-DYLE first generates a summary, and the EE model extracts evidence sentences for each generated summary sentence by computing their relevance scores and applying filtering methods, as described in Sections 5 and 6.2.

As illustrated examples, Tables 12 and 13 present the extracted evidence sentences and their extraction performances when using the Multi-DYLE's extractors and the EE model, given the query in Table 8. Note that CESs are used for gold evidence sentences in Table 12, whereas a union of CESs and PESs is used for gold ones in Table 13, and thus these scores are not fairly comparable. Given the differences in the evaluation settings, the extraction score in Table 12 is considerably higher than that in Table 13.

In Table 12, the ROG-based and CES-based extractors refer to those induced from Multi-DYLE(X_o^{ROG}, X_o^{CES}). In Table 13, the row 'Generated Summary' presents a generated summary, and the row 'Extracted Explainable Evidence' describes the evidence sentences extracted by the EE model for each sentence, P-Sent-n, in the generated summary.

The performances in Table 13 are *per-query* scores of the meeting summarization and E3 task, which is computed specifically for the query in Table 8; 'Summarization performance' indicates the per-query ROUGE score under the standard summarization metric, and 'E3 performance' refers to the per-query extraction ("summary-level") and Joint ROUGE scores as the joint evaluation metric, as in Section 6.3 and Table 5.

D Implementation Details

The proposed framework is implemented by extending the base DYLE model. To train Multi-DYLE, we explored different models for parameter initialization and used *RoBERTa-base* and *DYLE(generator)* to initialize the extractor and generator modules, respectively, as they show reliable performance in Table 6. To train the EE model, we fine-tuned the *RoBERTa-base* model with an Adam optimizer, learning rate of 5e-5, batch size of 8, and a gradient accumulation step of 8. We also utilized the ROUGE package to evaluate the performance of summarization and the NLTK library to preprocess the dataset.

We used an RTX 6000 NIVIDA GPU with a 48GB memory capacity, implemented and trained all models using the Pytorch library. To ensure the reliability of our results, we performed five distinct experimental runs, each with a different random seed, and stored the checkpoints with the maximum evaluation score.

Multi-DYLE	(Initialization Models) **Extractor**	**Generator**	**R-1**/ R-2/ R-L
(X_o^{ROG}, X_o^{CES})	*RoBERTa-base*	*BART-large*	37.17/ 12.26/ 32.75
	DYLE(extractor)	*BART-large*	36.54/ 11.65/ 31.92
	© *RoBERTa-base*	*DYLE(generator)*	37.55/ 12.43/ 32.76
	DYLE(extractor)	*DYLE(generator)*	36.97/ 12.03/ 32.20

Table 6: Performance of Multi-DYLE(X_o^{ROG}, X_o^{CES}) varying parameter initialization for extractor and generator modules on the meeting summarization task in the QMSum dataset. The DYLE's extractor and generator, which are publicly available by the authors of DYLE, are used for parameter initialization.

Type	Characteristic	Evidence Sentences	
Central Evidence Sentence (CES)	(1) exact	[G-Sent]	They also decided to start with basic functions and then move on to the more advanced feature.
		CES	Marketing : Well , should we start with just the core , the basic functions that we need .
		CES	Marketing : And then we can move on to the more advanced features .
	(2) semantic	[G-Sent]	The project manager briefed the team on some new requirements and initiated a discussion in which the team discussed and decided on various features to include in the remote they will produce.
		CES	Project Manager : So um I have to inform you I receive an email from the management bon board today and they have new requirements for the for the remote control . Um
	(3) supportive	[G-Sent]	The project manager opened the meeting and then the marketing expert discussed user requirements.
		CES	Marketing : Okay , so basically I'm gonna present some findings of a study we conducted uh into uh what users want in this remote control .
Peripheral Evidence Sentence (PES)	(1) illustrative	[G-Sent]	The project manager briefed the team on some new requirements and initiated a discussion in which the team discussed and decided on various features to include in the remote they will produce.
		CES	Project Manager : So um I have to inform you I receive an email from the management bon board today and they have new requirements for the for the remote control . Um
		PES	Project Manager : first um , they say that's uh about something about t teletext .
		PES	Project Manager : Um the second thing is uh they suggest that that we should uh use the remote control only for T_V_ , not for D_V_D_ and other devices ,
		PES	Project Manager : the third one is uh about the the the image of the company .
	(2) introductory	[G-Sent]	The remote will have buttons for channel changing, volume settings, numerals, and power on/off.
		PES	Project Manager : but first maybe what is what are the usual function of a standard remote control ?
		CES	Marketing : Okay , well , I mean the obvious one is changing channels .
		CES	Project Manager : So , turning channel , of course . Volume setting .
	(3) connective	[G-Sent]	Whether using radio waves will interfere with other technology a user owns.
		CES	Marketing : Do you think radio waves um will interfere with other appliances in the home ?
		PES	User Interface : Uh , I don't think so ,
		PES	User Interface : because uh we can make uh we ca we can make this wave in a specific frequency .
		CES	User Interface : So they can be in a range which is not inter interfering with the with other devices inside the home .

Table 7: Examples of Central Evidence Sentence (CES) and Peripheral Evidence Sentence (PES). Each row shows CESs or PESs aligned with each gold summary sentence referred to by [G-Sent].

Query	Summarize the discussion about trend watching and appearance design.
Gold Summary	(G-Sent-1) The marketing put forward three noteworthy aspects in trends. (G-Sent-2) First and foremost, people loved fancy things that they could be identified with. (G-Sent-3) The second point was that as a remote control it had to be technologically innovative. (G-Sent-4) Thirdly, being easy to use was also necessary. (G-Sent-5) From a broader perspective, fruit and vegetables were in fashion this year and being spongy was also popular. (G-Sent-6) Thus, contrary to the industrial designer, the marketing thought rubber was more feasible in terms of sponginess. (G-Sent-7) The group agreed that the product should resemble fruit and vegetable in shape and colour but the specific design was not decided.

	Gold Summary	**Type**/subtype	Evidence Sentences	Turn-Sent Index
Evidence Alignment	G-Sent-1	**PES**/intro.	Marketing : I'm going to talk about trends	399-5
		CES/sem.	Marketing : but uh basically there are uh in in the market of of remote controls there are three aspects that we should very pay much attention to .	403-4
	G-Sent-2	**CES**/sem.	Marketing : The first one , which seems to be the most important one , is that it has to be fancy , it has to have a fancy look and feel .	403-5
	G-Sent-3	**CES**/exact	Marketing : Strangely enough it's more important to be fancy than to be wi and now that's the second thing it has to be , it has to be technologically i innovative ,	403-7
	G-Sent-4	**CES**/sem.	Marketing : which is that it should be easy to use and it should be easy to use as a remote control .	405-3
	G-Sent-5	**CES**/sem.	Marketing : Uh and now in a more uh general uh uh broad way of seeing th uh the thing .	410-0
		CES/sem.	Marketing : currently the the trends that we see in l in l big cities like Paris and Milan , well , it seems that this year things should have uh a fruit and vegetable uh way of of look or feel	410-3
		CES/sem.	Marketing : And uh if we co we compare to last year , now it has to be spongy ,	417-1
	G-Sent-6	**PES**/intro.	Marketing : When we were talking about rubber ,	425-0
		PES/conn.	Marketing : I think uh the rubber aspect might be important	427-0
		CES/exact	Marketing : because it's what is probably more feasible in terms of sponginess .	427-1
	G-Sent-7	**CES**/sem.	Marketing : We have to I think we have to have the look of fruit and vegetables .	477-0
		CES/supp.	Industrial Designer : fruit . These things can be easily incorporated .	485-4
		CES/sem.	Industrial Designer : We can have t colours or this shape	485-5
		CES/supp.	Project Manager : Now we have to decide on what kind of fanciness .	551-0

Table 8: Example in ExplainMeetSum, with a full set of CESs and PESs aligned with a summary for the query '*Summarize the discussion about trend watching and appearance design*" in QMSum (i.e., the transcript name: IS1006c, the query id: specific-6); In the upper part, a gold summary is provided where G-Sent-n refers to n-th gold summary sentence; In the lower part, a set of CESs and PESs aligned for each G-Sent-n are presented.

Tag Refinement Before	Tag Refinement After		Examples
CES/ supportive	CES/ semantic	Similarity Score	(SBERT) 72.20
		G-Sent	A **curriculum reform** was to **carry out** throughout **Wales**.
		Evidence Sentence	Meilyr Rowlands : So, we need to ensure that those qualifications are reformed as a result of the **reform** of the **curriculum**, and, of course, Qualifications **Wales** is **carrying out** that work currently.
PES/ introductive	CES/ supportive	Similarity Score	(SBERT) 53.53
		G-Sent	When it comes to continuing **mental health** service during the lockdown, **Vaughan Gething** insisted that it **was** of great necessity to carry out a **mental health** recovery plan that with such a system, government can ensure the **children** could enjoy a healthy **mental** state during the school lockdown.
		Evidence Sentence	**Vaughan Gething** AM : So , **children 's mental health** was a central concern and remains so for both myself and the education Minister .
CES/ supportive	PES/ connective	Similarity Score	(SBERT) 42.69
		G-Sent	When discussing the governmental issue of dealing with systematic racism, Justin Trudeau mentioned that actually there had been serious systematic racism in most national institutions for the past two years, so he called for a **revolution in those organizations** to welcome equal cooperation with the black colleagues and indigenous communities.
		Evidence Sentence	Mr. Jagmeet Singh : Is the Prime Minister committed to a full-scale overhaul of the RCMP to root out systemic **racism?**

Table 9: Query control method via the semi-automatic test suite based on neural similarities between evidence and summary sentences. The 'Tag Refinement' column presents how the initial erroneous labels on sample evidence sentences are revised correctly after checking their neural similarities with gold summary sentences.

	#	Turn Index	Evidence Sentences
Turn -level Oracle*	1	4	User Interface : (4-0) How was lunch ?
	2	140	Project Manager : (140-0) Three .
	3	168	User Interface : (168-0) I thought you like it . (168-1) Ah okay
	4	278	Marketing : (278-0) The the young people the young people want to be different from their friends .
	5	367	Marketing : (367-0) Okay . (367-1) So it could be smart in that way .
	6	392	Project Manager : (392-0) yeah , Marketing Expert .
	7	405	Marketing : (405-0) it has to be new with some of uh new uh technology inside (405-1) and uh and this is also uh more important than the last thing (405-2) which we w may think that would have been the most important , **(405-3) which is that it should be easy to use and it should be easy to use as a remote control .** (405-4) So as you see uh it first have to be very nice , (405-5) s something that people are proud of (405-6) uh uh that i uh they can be id identified with (405-7) uh and and then uh something that um contains very novel stuff (405-8) that they can talk about with their friends , huh , mine has this and not yours . (405-9) And finally of course it has to be useful as a remote control (405-10) but it seems that it's not so important that it's useful as a remote control .
	8	425	Marketing : **(425-0) When we were talking about rubber ,**
	9	427	Marketing : **(427-0) I think uh the rubber aspect might be important (427-1) because it's what is probably more feasible in terms of sponginess .**
	10	438	Marketing : (438-0) Think more of uh something in the colours of uh like fruit and vegetables and spongy ,
	11	439	Industrial Designer : (439-0) Fruit . (439-1) Even shape ?
	12	595	Industrial Designer : (595-0) Even design .

* Turn-level Oracle (Mao et al., 2022)

Table 10: Example of a *turn-level* ROUGE-based extractive oracle for a gold summary in Table 8 the bold-faced numbers refer to the turn-based sentence ids of the annotated CESs or PESs.

	Example Performance	#	Evidence Sentences	Turn-Sent Index
Sentence -level Oracle	(P) 15.38 (R) 16.67 (F1) 16.00	1	Project Manager : yeah , Marketing Expert .	392-0
		2	Marketing : but uh it's not so simple .	399-14
		3	**Marketing : which is that it should be easy to use and it should be easy to use as a remote control .**	**405-3**
		4	Marketing : uh uh that i uh they can be id identified with	405-6
		5	Marketing : And finally of course it has to be useful as a remote control	405-9
		6	Marketing : That's the thing with trends	415-1
		7	Marketing : Fruit and vegetable . Think fruit and vegetable .	417-0
		8	**Marketing : because it's what is probably more feasible in terms of sponginess .**	**427-1**
		9	Marketing : Think more of uh something in the colours of uh like fruit and vegetables and spongy ,	438-0
		10	Industrial Designer : that	483-0
		11	Marketing : it has to be fancy	541-3
		12	Industrial Designer : Even design .	595-0
		13	Project Manager : explore a shape .	603-1

Table 11: Example of a *sentence-level* ROUGE-based extractive oracle for a gold summary in Table 8; the bold-faced numbers refer to the turn-based sentence ids of the annotated CESs or PESs.

	Turn-Sent Index	Evidence Sentences	Summarization Extraction Performance
ROG-based Extractor	541-3	Marketing : it has to be fancy	
	513-0	Industrial Designer : we want to follow general trend .	
	399-10	Marketing : first maybe just a small recap on how how do we watch trends	
	548-0	Marketing : It's fancy .	
	403-5	**Marketing : The first one , which seems to be the most important one , is that it has to be fancy , it has to have a fancy look and feel .**	
	468-0	Industrial Designer : it's not particular to the remote control .	
	425-0	Marketing : When we were talking about rubber ,	(P) 6.67
	533-0	Project Manager : So titanium smell like fruit .	(R) 8.33
	543-0	Industrial Designer : Feature	(F1) 7.41
	399-5	Marketing : I'm going to talk about trends	
	444-0	And not those futuristic uh remote control with angles and uh and titanium like .	
	451-0	Marketing : but that's that's fashion	
	466-0	Industrial Designer : It's more general trend	
	458-2	Industrial Designer : Or it's	
	553-1	Industrial Designer : we will try to explore these two options	

	Turn-Sent Index	Evidence Sentences	Summarization Extraction Performance
CES-based Extractor	513-0	Industrial Designer : we want to follow general trend .	
	461-1	Marketing : we have people uh uh listening to the trends everywhere in the world , of course ,	
	509-3	Industrial Designer : and we want some themes like fruits or vegetables ,	
	430-1	Project Manager : So maybe titanium it's not a good idea .	
	509-2	Industrial Designer : and we want some kind of buttons	
	278-0	Marketing : The the young people the young people want to be different from their friends .	
	509-1	Industrial Designer : we want the speech recogniser	
	403-5	**Marketing : The first one , which seems to be the most important one , is that it has to be fancy , it has to have a fancy look and feel .**	
	438-0	Marketing : Think more of uh something in the colours of uh like fruit and vegetables and spongy ,	(P) 26.67
	463-1	Marketing : and uh so I'm just asking them what are the current trends according to them when they go in the stores and when they ask uh their uh friends	(R) 33.33 (F1) 29.63
	403-4	**Marketing : but uh basically there are uh in in the market of of remote controls there are three aspects that we should very pay much attention to .**	
	304-0	Marketing : It has this distinctive look and feel and look	
	410-3	**Marketing : currently the the trends that we see in l in l big cities like Paris and Milan , well , it seems that this year things should have uh a fruit and vegetable uh way of of look or feel**	
	477-0	**Marketing : We have to I think we have to have the look of fruit and vegetables .**	
	288-1	User Interface : But you know if you want to be different you just take your remote control with you all the time .	

Table 12: Evidence sentences predicted by Multi-DYLE's extractors and their extraction performances for the query in Table 8 where CESs are used as gold evidence ones; ROG-based and CES-based extractors are ones induced from Multi-DYLE(X_o^{ROG}, X_o^{CES}) in Section 6.1.1.

	Predicted Summary	Summarization Performance
Generated Summary	(P-Sent-1) according to marketing , there were three aspects in the market of remote controls that the team should pay much attention to .	
	(P-Sent-2) first , the fancy look and feel was the most important .	ROUGE
	(P-Sent-3) young people wanted to be different from their friends , so they should take their remote control with them all the time .	(R-1) 45.83
	(P-Sent-4) the second was to follow the general trend of fruits and vegetables , which was seen in big cities like milan and paris this year .	(R-2) 10.53 (R-L) 43.75
	(P-Sent-5) then , industrial designer proposed to explore the two options of speech recogniser and buttons .	

	Predicted Summary	Turn-Sent Index	Evidence Sentences	E3 Performance
Extracted Explainable Evidence	P-Sent-1	**403-4**	**Marketing : but uh basically there are uh in in the market of of remote controls there are three aspects that we should very pay much attention to .**	
		405-9	Marketing : And finally of course it has to be useful as a remote control	
	P-Sent-2	**403-5**	**Marketing : The first one , which seems to be the most important one , is that it has to be fancy , it has to have a fancy look and feel .**	Summary -level
	P-Sent-3	278-0	Marketing : The the young people the young people want to be different from their friends .	(P) 25.00
		288-1	User Interface : But you know if you want to be different you just take your remote control with you all the time .	(R) 20.00 (F1) 22.22
	P-Sent-4	**410-3**	**Marketing : currently the the trends that we see in l in l big cities like Paris and Milan , well , it seems that this year things should have uh a fruit and vegetable uh way of of look or feel**	Joint ROUGE
		509-3	Industrial Designer : and we want some themes like fruits or vegetables ,	(R-1) 58.65
		513-0	Industrial Designer : we want to follow general trend .	(R-2) 36.20
		509-1	Industrial Designer : we want the speech recogniser	(R-L) 55.64
	P-Sent-5	509-2	Industrial Designer : and we want some kind of buttons	
		553-1	Industrial Designer : we will try to explore these two options	
		556-0	Marketing : Maybe you could explore the two option .	

Table 13: Full-pipelined examples of the meeting summarization and E3 tasks with their summarization/extraction performances for the query in Table 8. A summary is generated by Multi-DYLE(X_o^{ROG}, X_o^{CES}) (i.e., $M = 2$), and the evidence sentences are extracted by the EE model. Unlike Table 12, a union of CESs and PESs is used as gold evidence for computing the extraction score. The Joint ROUGE score defined in Section 6.3 is also provided.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, it is discussed in Limitations Section.

☒ A2. Did you discuss any potential risks of your work?
No, there are no potential risks of our work, as our work made additional annotation in the widely-used QMSum dataset and addressed the meeting summarization and evidence extraction tasks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, the paper's main claim is stated in Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
No, we have not used AI writing assistants.

B ☑ Did you use or create scientific artifacts?

Yes, the artifacts for the dataset are discussed in Section 3, with its URL in Abstract, and the model part is described in Sections 4 and 5.

☑ B1. Did you cite the creators of artifacts you used?
Yes, we have cited the creators of artifacts in Section 1.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No, we have cited the artifacts, but did not explicitly discuss the license or terms for use as it is understood in the NLP field to share or utilize related artifacts. For our dataset, we will specify the license name in the URL, when available.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, we have repeatedly discussed and specified the intended use of existing artifacts for our artifacts in Sections 1, 2, 3 and 4. The details of the datasets we used are presented in Section 3.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No, we do not reveal any new contents in our artifacts, because our annotation was made on the contents in the QMSum dataset. Thus, there is no new critical information in our dataset, such as individual people and names.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, we have provided basic information about the data in Section 3.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, we have provided basic information about the existing and created data in Section 3.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Yes, we have described results of various computational experiments in Section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, we have provided implementation details in Appendix D.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, the experimental setup for the best-performing model, including the used hyperparameters for the Multi-DYLE and Evidence Extraction (EE) model, are presented in Tables 2-4 in Section 6. The additional details on the setup are provided in Appendix D.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes, we have provided the basic statistics such as the number of runs tried in Appendix D.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes, we have provided it in Appendix D.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes, we used human annotators to build our dataset, as in Section 3 and Appendix A.1.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Yes, but instead of providing separate instructions for annotators, we established the feedback-based annotation protocol, leading by a coordinator and an expert, which are frequently communicated with the annotators, as in Appendix A.1 and A.3.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Yes, we have provided the detailed annotation process in Appendix A.1.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No, the annotators we recruited only performed the labeling work on the QMSum. No new text which require agreement was produced during the process.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No, the necessary verification has already been performed at the source of the data. We directly use the original QMSum dataset without making any changes to it.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No, the necessary verification has already been reported at the source of the data.