Context or No Context? A preliminary exploration of human-in-the-loop approach for Incremental Temporal Summarization in meetings

Anonymous EMNLP submission

Abstract

Incremental meeting temporal summarization, summarizing relevant information of partial multi-party meeting dialogue, is emerging as the next challenge in summarization research. Here we examine the extent to which human abstractive summaries of the preceding increments (context) can be combined with extractive meeting dialogue to generate abstractive summaries. We find that previous context improves ROUGE scores. Our findings further suggest that contexts begin to outweigh the dialogue. Using keyphrase extraction and semantic role labeling (SRL), we find that SRL captures relevant information without overwhelming the the model architecture. By compressing the previous contexts by $\approx$ 70%, we achieve better ROUGE scores over our baseline models. Collectively, these results suggest that context matters, as does the way in which context is presented to the model.

1 Introduction

In meetings, especially in a virtual setting, distractions are common place and can last anywhere from a few seconds to minutes, impacting concentration and participation in the remainder of the meeting negatively. A note-taking tool designed to provide temporally relevant summaries of what has happened in the last 2-3 minutes may mitigate the negative effects of distractions and interruptions.

Missing a few minutes of content, rather than the whole meeting, provides unique challenges for current summarization tools. Instead of summarizing the main points of the meeting, a temporally-relevant summarization aid must instead capture relevant meeting content given previous events, even if those events would not be included in the full meeting summary. Such a tool may benefit from taking the past notes or summaries from meeting participants as context and incrementally updating the summaries for a specific time interval to capture relevant information that a distracted individual would need to know to reintegrate into the meeting.

The goal of this work is to investigate the ability to incrementally summarize meetings, specifically focusing on how a summarization tool may make use of past summaries to increase the accuracy of temporally-relevant abstractive summarization.

The task of incremental temporal summarization in dialogue has two main aspects to it, i) The content being summarized has a temporal order–the information evolves over time. ii) summaries build upon or use the past context (transcriptions, summaries, or human notes) to generate the summaries for the current dialogue. A new dataset based on incremental temporal summarization of the AMI dataset, which we call the AMI-ITS, provides a means to investigate incremental temporal summarization of meeting dialogues.

Temporal summarization has been studied in the context of summarizing news articles (Dang and Owczarzak, 2008; McCreadie et al., 2014; Aslam et al., 2015). In such a setting, the input news articles that evolve over time are streamed in chunks. The summarizer needs to either summarize the new content or update the earlier generated summary with the new information. While similar to incremental temporal summarization (ITS) in meetings scenario, additional challenges are associated with the properties of human conversation such as disfluencies and dyadic exchanges (questions and answers, acknowledgements, confirmations etc.) where a contributions to the summaries are from multiple interlocutors (Poesio and Rieser, 2010).The information also comes in smaller increments of time, and at a much faster rate than news articles. Limited work has been done on temporal summarization and incremental summarization in multi-party meeting scenarios.

The main contribution of this work is to quantify the impact of previous human generated summaries in improving meeting summarization. We specifically focus on how to best use previous summaries from earlier temporal summarization. This mimics the use of the meeting notes of individuals to generate up to date summaries of meeting dialogue and provides the basis for an incremental summarization tool that works jointly with meeting participants in real time. We ask fundamental questions about how to use previous summaries by humans including whether meeting summaries or meeting dialogues should be prioritized as input to the model. We then look at how many summaries the model requires to most accurately summarize the most recent temporal chunks and conclude by showing that extracting meaningful information from past summaries through semantic role labeling can further improve temporal summarization. Collectively this work shows that temporal summarization benefits from having a human in the loop and suggests ways to use human input most effectively.

2 Related work

Because of the differences between news articles and human dialogue, incremental summarization for meetings/dialogues provides unique challenges and requires novel approaches. Table 1 compares training examples and summarizations across a standard news corpus (CNN/DailyMail), scientific paper summarization (Pubmed), the AMI meeting corpus, and the temporal version of the AMI meeting corpus (AMI-ITS) which focuses on 100 second incremental temporal sequences from the AMI dataset and will be explained in more detail below. Not only are the meeting corpora much smaller in terms of training examples, the dialogue is much longer compared to news articles, averaging 4757 words in the AMI meeting transcripts compared to 781 words for the news corpus. While meetings tend to be much longer in length than news articles, much of this information is considered non-extractive (i.e. not containing information relevant to the abstract summary). Incremental summarization is a noticeably different task than full meeting summarization, news summarization, and article summarization, with most of the words spoken being labeled as extractive. The summaries in the AMI-TS dataset are also longer than either the news corpus or the AMI corpus and the summaries are more than 25% of the overall extractive text. The novel challenge in temporal summarization for meeting dialogues is that much of the meeting text is relevant in summarizing key events and concepts of the previous 100 second chunks. These differences suggest that the temporal summarization task is different from news summarization and full meeting summarization in two main ways 1) meetings have different properties than other types of text and 2) temporal summarization is different than summarizing a whole document.

Corpus	doc.	obs.	words	extract	summary (%)
CNN/DM	312K	312K	781	382	56 (7.2%)
Pubmed	133K	278K	3016	-	203 (6.7%)
AMI	137	137	4,757	210	19 (0.4%)
AMI-ITS	49	924	262	162	67 (25.6%)

Table 1: Corpus statistics: number of documents, number examples, average number of words, proportion of extractives and the average number of words in the abstractive summary for each example.

Meeting Summarization. Much of the available summarization datasets exist for news articles summarization scenario (Narayan et al., 2018; Dernoncourt et al., 2018). The news articles and summaries for these news articles have a very different structure than meetings and dialogue. Dialogue summarization corpora (Carletta et al., 2005; Janin et al., 2003; Lacson et al., 2006; Favre et al., 2015; Misra et al., 2015; Barker et al., 2016; Liu et al., 2019a; Gliwa et al., 2019) have helped accelerate the research in the area of conversational summarization. Major differences exist between dialogue summarization and summarization of news articles (Jung et al., 2019). News articles tend to follow a structure in which the most relevant information is contained early in the text. Meetings, by definition, require engagement of multiple participants resulting in transcripts with different styles, perspectives, and roles. Compared to news summarization, labeled training data of meeting summaries is also severely limited. Several models have been developed recently focused on generating summaries for meetings and dialogues and have achieved promising results (See for e.g. See et al. (2017); Chen and Bansal (2018); Zhao et al. (2019); Liu (2019); Zhang et al. (2020); Feng et al. (2020); Zhu et al. (2020); Fabbri et al. (2021b)). These models suggest that altering the input representation, the model architecture and loss function may all play a part in improving accuracy for summarization of meetings.

Incremental Summarization. While meeting summaries are limited by datasets, incremental tem-

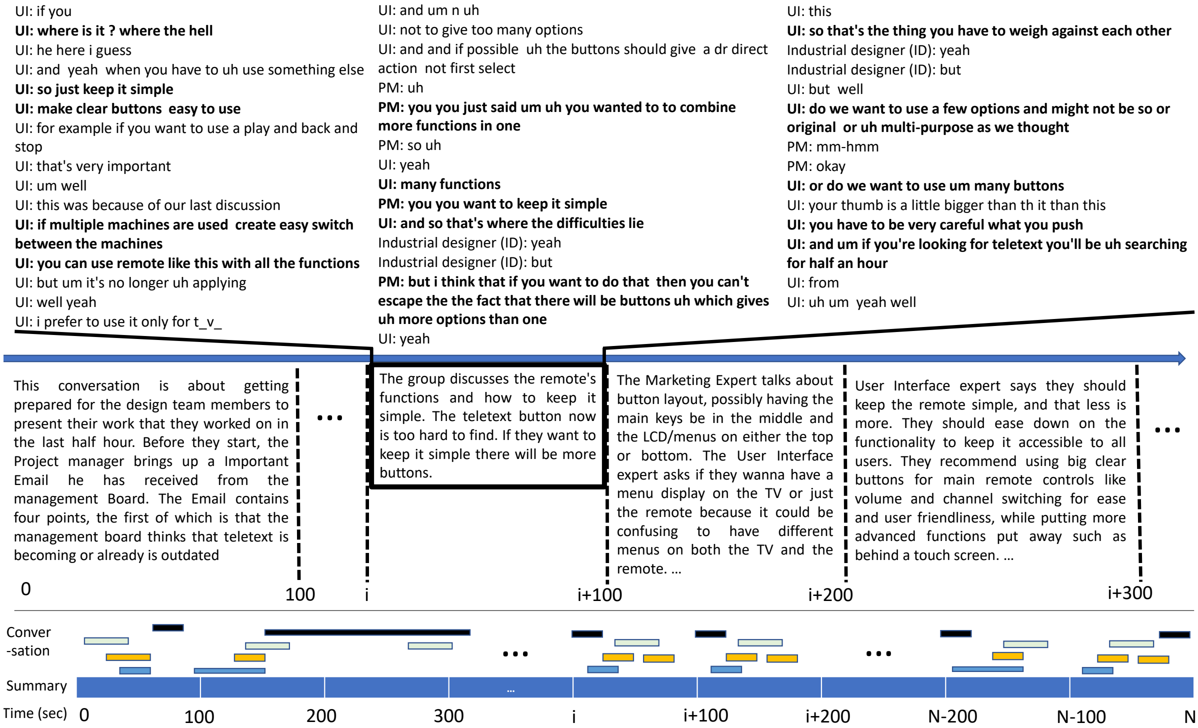

Figure 1: Shows sample incremental temporal summaries from the corpus along with the conversation transcriptions and extractives (in bold) as marked by a crowd-worker.

poral summarizations of meetings is even more limited. Instead of focusing on summarizing the full content of the meeting/dialogue, incremental summarization focuses on building incremental representations of the meetings rather than a full summary at the end. Work in incremental dialogue processing has considered when to add additional information to an existing summary (McCreadie et al., 2014), how representations of individuals and topics can be influenced by time (Chen and Metze, 2012), considerations of turn taking (Zhu et al., 2020) and more (Zhong et al., 2021). While these models consider various aspects of incremental and temporal summarization in the model design choices, evaluation often excludes incremental and temporal aspects.

Recently, deep learning models (Li et al., 2019; Liu et al., 2019b) and especially transformer-based models, have achieved impressive performance in abstractive summarization task (Zhang et al., 2020; Raffel et al., 2020; Lewis et al., 2020; Zhu et al., 2020). Such transformer-based models are typically pre-trained on a large dataset and then fine-tuned on a smaller dataset. In this work, we adopt a current state-of-the-art transformer architecture, BART, and utilize and evaluate transfer learning to generate temporally relevant summaries to meeting dialogue. Recent work focusing on meeting summarization has suggested that a new architecture (HMNet) may improve summarization on meeting dialogue (Zhu et al., 2020). This work extends transformer architectures to include a word level transformer, to process and encode the word-level dialogue, and a turn-based transformer which considers the speaker role and sentence embeddings from the word-level transformer. This model architecture has achieved SOTA performance on the AMI meeting corpus but has not been validated on incremental summarization tasks. Our contribution is not to develop a new model architecture for summarization or to outperform state-of-the-art but rather to examine the role of previous summaries on the ability to improve performance in later summaries. We hope to understand the usefulness of previous summaries (contexts) in accurately summarizing the current temporal information. We leave temporal summarization using such architectures to future work.

3 Data

Our primary focus is on abstractive summarization for incremental temporal scenarios. The incremental temporal summarization module takes the utterances in the current time window as input. In this work, we focus on how best to use the past summaries (context) as input. Models are evaluated on temporal summaries capturing the last 100 seconds of the meeting. While it has been shown that having previous temporal summaries is helpful in accurately summarizing a specific context (Manuvinakurike et al., 2021), we investigate this question further by asking how much context is relevant and how to best use past context. We use these results to draw conclusions about the role of human summarization in model performance.

AMI/AMI-ITS corpus: In this work, we rely heavily on a novel extension to the AMI-meeting dataset (Carletta et al., 2005) which we call the **AMI-ITS** dataset (Manuvinakurike et al., 2021). The meetings in the original AMI dataset consist of conversations between 4 role-playing participants (Project Manager (PM), Industrial Designer (ID), User Interface expert (UI), and Marketing expert (ME)) in a remote-control design scenario. Each group of 4 participants meet 4 times and continue the conversation forward from the previous sessions but often on a new agenda. The AMI corpus consists of extractive and abstractive summaries for the full conversation annotated by experts.

The AMI-ITS dataset provides extractive and abstractive summaries for 100 second time durations on a subset of the AMI meetings. Table 1 indicates the number of 100 second chunks that were labeled in the AMI-ITS corpus and the average number of tokens in the full text, extractive and abstractive summaries. We refer to the original AMI dataset, specifically the extractive and abstractive summaries, as AMI and use the addition of ITS to indicate the *incremental and temporal* meeting dialog corpus. To build the AMI-TS corpus, individuals were presented with a 100 second dialogue chunk. They also saw up to 3 summaries that captured the 3 preceding dialogue chunks. Participants would check a box next to each line of text indicating whether or not the specific dialogue line was extractive, or relevant to the summary. They then provided a summary of the dialogue which was used as context for down-stream meeting dialogues. Figure 1 shows a sample incremental temporal summary from the AMI-ITS dataset.

We evaluate all models on their ability to predict abstractive summaries from AMI-ITS. In all cases, 3 models of each type were trained to compute average performance and estimate model variability. We select models to optimize ROUGE-1 recall values but also report other measures. In total 42*3 models were trained for this work.

4 Models

Model Input/Output: The input to all models is extractive meeting dialog. For this work, we use human judgements of extractive sentences as labeled by participants in the AMI-ITS data collection pipeline. Previous work by Manuvinakurike et al. (2021) showed that learning a highly accurate automatic extractor given available training data is possible with accuracy above 70%. Role information ($role$, e.g. 'Project Manager (PM):') may be included as part of the input as well. Work on dialog summarization indicates that role information is important in abstractive summarizations (Zhu et al., 2020) and thus we include comparisons of role and non-role labeled dialogues in our experiments.

Context: The main model variants investigate the role of context in improving abstractive summarization. We define context to be the number of previous (human generated) abstractive summaries provided to the model during training and prediction. For example, our summarization model may be asked to summarize the meeting events that happened between 1000 and 1100 seconds of a given meeting. In this case, there are 10 previous contexts that the model can be provided. Because the temporal summaries are focused on only the events of 1000 to 1100 seconds, the summarization model may not benefit from seeing summaries from the first 0 to 100 seconds but may benefit from seeing the summary from 900 to 1000 seconds.

In labeling our models and results, we include the number of past summaries the model saw during training. A context value of 0 indicates that the summarization model was provided no summaries from the past, whereas, a context value of 5 would indicate that summaries for the most recent 5 100-second chunks were included. Because of the redundancy in the transformer model input as context values increase in length, the order of the previous contexts is shuffled. Each context is separated by the end of sentence, start of sentence characters from the model tokenizer.

5 Methods and Results

We focus our exploration on BART as the baseline model as this model has been investigated both in incremental summarization and dialogue summarization. For fine-tuning of abstractive models, we fine-tune for a maximum of 25 epochs and choose the model resulting in the best ROUGE-1 F-measure on the validation set. We use the following configuration for all baseline models: learning rate=0.0001, training batch size=4, and label smoothed negative log-likelihood loss. The maximum sequence length is set to 1024. The models can generate summaries of the max length of 142 tokens. For model training and inference, we use multiple machines with a combination of either an Intel(R) Xeon(R) or Intel(R) Xeon(R) Platinum 8280 CPU and NVIDIA Titan X or Titan Xp GPU. All models were trained on 2 GPUs. For the pretrained models, we use the BART-large-cnn model, from the Huggingface (Wolf et al., 2019) library. We retain the default model configurations. For all experimental conditions, we input the transcriptions of the extractives marked by crowd workers in the AMI-ITS dataset and n previous contexts. The order of the previous contexts are randomly shuffled when building the dataset. We evaluate models on their ability to generate the abstractive summaries similar to those provided by the crowd workers in the AMI-ITS dataset.

5.1 Fine-tuning to dialogue

We first investigate whether incremental temporal summarization is improved by fine-tuning a pretrained summarization model, originally trained on CNN/DailyMail (CNN), to meeting dialogues and their respective abstractive summaries from the AMI corpus. As mentioned, news summarization often emphasize and leverages information from early in the news article; dialogue does not follow any systematic structure and the beginning of meetings may actually contain spurious information such as introductions and technical issues.

Because the task is to summarize small chunks of dialogue, it is possible that the granularity of the AMI summaries, which is significantly less than required for 100 second time slices, not improve the performance over the baseline model. Thus we compare using the pretrained BART-large model, trained on CNN news articles (Hermann et al., 2015; Nallapati et al., 2016) to one that is fine-tuned on the AMI dataset (Carletta et al., 2005) (AMI). In all cases, we fine-tune on the training data portion of the AMI-ITS dataset and evaluate on the AMI-ITS test set. We also consider the importance of speaker role information by using role labels in the AMI dataset and role labels at test. and fine-tuning both models on AMI-ITS dialog that contains role information. Baseline models are evaluated by ROUGE scores (R1, R2 and RL)[1] on a testing set of the AMI-ITS dataset.

We conclude from table 2 that fine-tuning on the AMI dataset may hurt performance on the AMI-ITS dataset. It is unclear if role information affects performance. The decrease in performance when fine-tuning on AMI is likely due to the difference in tasks–summarization of a full meeting versus summarization of the last 100 seconds. We thus use the pretrained BART CNN transformer for all subsequent experiments.

model	ROUGE-1	ROUGE-2	ROUGE-L
CNN	47.61/34.14	**15.28/11.21**	29.07/**20.36**
CNN_{role}	**47.85**/33.80	**15.47**/11.01	**29.17**/20.07
AMI	45.27/**35.42**	14.38/11.14	28.16/21.34
AMI_{role}	45.71/33.85	13.89/10.15	27.89/20.10

Table 2: R1, R2, and RL scores (recall/precision) on the AMI-TS dataset for BART trained on CNN/DailyMail (CNN) or fine-tuned first on AMI (AMI). $role$ indicates speaker role information is part of the input.

5.2 Summaries vs extractive texts

As we add more and more previous contextual information to the model, the input length quickly exceeds the max length that the pretrained model can process. In the case of the BART CNN/DailyMail model, inputs larger than 1024 tokens are ignored. This can be problematic when training and evaluating performance of the BART AMI-TS model specifically because the model may be using the text and summary information differently. We thus ask whether model performance changes when we truncate the input, preferring to maintain either 1) extractive text information or 2) context information. To investigate this question we consider input representations that include extractive text and up to 10 previous summaries where available. We then test two model variants: one that will maintain the extractive text to the exclusion of the summaries and another than maintains the summaries to the exclusion of the extractive text. Table 3 shows that model performance is positively affected by the availability of the extractive text than models preferring previous summaries over current text in terms of R1 recall. This highlights a difference between human summarization and model summarization as Manuvinakurike et al. (2021) showed

[1]R0UGE scores were calculated via rouge-score version 0.0.4 pypi.org/project/rouge-score/

model	ROUGE-1	ROUGE-2	ROUGE-L
T-10	**46.11**/34.39	13.60/10.21	**27.92**/20.37
T-10$_{role}$	45.35/34.47	13.90/10.78	**27.92**/20.62
C-10	44.23/**36.37**	14.25/**12.12**	27.47/**22.21**
C-10$_{role}$	44.32/36.12	**14.27**/11.66	27.80/22.00

Table 3: R1, R2, and RL (recall/precision) scores for models that selectively prefer extractive text over contexts (T-10) or contexts over extractive text (C-10) in the case wher 10 contexts are used.

that human summaries were higher quality when previous contexts were supplied. For the rest of our experiments, we keep extractive text over summaries when the input length exceeds the maximum length of the model input.

5.3 The effect of past summaries

Our main research question focuses on to what extent previous (human) generated summaries improve the quality of the summaries. To explore this question, we construct model inputs that include a various number of previous temporal summaries. We consider models trained without and with role labels on the dialogue. Table 4 shows the result from this experiment. Generally, the quality of the summaries from a model trained on input without the role information does not improve with the addition of summary information when evaluated on ROUGE recall. We see a small improvement in ROUGE precision. It may seem non-intuitive that additional contexts does not improve ROUGE recall, but this result may be because the model receives large amounts of context information compared to dialogue, resulting in over-attendance to past summaries rather than current dialogue.

In the case of a model trained with role labels on the dialogue, previous contextual information helps, up until a point. For improving recall, providing the previous 5 summaries improves performance and surpasses model performance when no role labels are provided. Precision is also highest when context information of 3 previous summaries is included as input to the model. These results suggest that previous context is useful to these models but that distinguishing contexts from dialogue is important to model performance.

5.4 Capturing context

Given the challenges of dealing with input length while including past contexts, we explore ways to capture only the relevant information from the past summaries. In this section we describe the methods for capturing the context using keyphrase extraction and semantic role labels from the past summaries.

context	ROUGE-1	ROUGE-2	ROUGE-L
0	**47.61**/34.14	**15.28**/11.21	**29.07**/20.36
1	46.88/34.82	14.05/10.54	28.72/20.59
3	45.89/**35.70**	14.93/**11.55**	28.36/**21.51**
5	46.81/34.55	13.87/10.13	28.50/20.50
10	45.35/34.50	13.93/10.80	27.90/20.62
0$_{role}$	47.85/33.80	15.47/11.01	29.17/20.07
1$_{role}$	46.22/35.50	14.14/10.90	28.25/21.08
3$_{role}$	45.34/**36.58**	14.33/**11.56**	27.70/**21.85**
5$_{role}$	**48.29**/33.67	**15.66**/10.85	**29.52**/19.88
10$_{role}$	46.65/34.52	14.28/10.54	28.35/20.44

Table 4: R1, R2, and RL scores (recall/precision) for models trained with different numbers of contexts.

Keyphrase extraction: For keyphrase extraction, we define the context as the 10 most important words or phrases from past summaries. To extract meaningful keyphrases from the human generated summaries, we use a pre-trained BERT model, KeyBERT (Grootendorst, 2020). This technique uses BERT-embeddings (Devlin et al., 2018) and cosine similarity to find sub-phrases in a document that are most similar to the full document itself. We generate top-10 keyphrases (ranging between 1-5 words) for each previous summary and use these keyphrases as past contexts. We use Maximal Margin Relevance (MMR, Carbonell and Goldstein (1998)) to reduce redundancy and increase diversity in the keyphrases. All keywords for each context are concatenated into one string and separated by end/start tokens. Results from table 5 indicate that keyphrase extraction improves ROUGE precision values but does not improve recall.

model	context	ROUGE-1	ROUGE-2	ROUGE-L
baseline	0	47.61/34.14	15.28/11.21	29.07/20.36
baseline$_{role}$	5	48.29/33.67	15.66/10.85	29.52/19.88
Keyphrase	1	44.57/**37.11**	13.35/11.23	26.85/**21.90**
	3	43.51/**36.81**	13.52/**11.61**	27.01/**22.35**
	5	46.61/34.70	14.39/10.86	28.53/20.75
	10	46.66/**35.33**	13.68/10.42	28.42/20.84
Keyphrase$_{role}$	1	46.54/**37.10**	14.96/**12.07**	28.01/**21.74**
	3	44.05/**37.58**	13.65/**11.74**	27.32/**22.84**
	5	46.92/**34.79**	15.52/**11.52**	**29.78**/21.45
	10	42.50/**36.97**	13.03/**11.47**	26.29/**22.26**

Table 5: R1, R2, and RL scores (recall/precison) for models trained with different amounts of past contexts where contexts are defined as the top 10 keyphrases extracted via keyBERT. Bolded values indicate improvement over baseline context models.

Semantic Role labeling: We next consider whether semantic role labels can provide relevant contextual information. Using semantic role labelers (SRL) for extracting semantic role information has shown promise, but remains largely unexplored (Yan and Wan, 2014; Trandabat, 2011). SRL helps extract important semantic information from the text in the form of Verb-Argument (& modifiers) which can serve as keywords to capture context. We extract semantic roles using Allennlp toolkit (Gardner et al., 2018) using a BERT-based model (Shi and Lin, 2019) trained on Ontonotes 5.0 dataset (Pradhan et al., 2013). The model is used out-of-the-box to extract verbs, and for each verb we also extract the verb arguments, including agents, patient, causers, instrument, benefactive, attribute, experiencers, starting point and ending points. These are ARG0-4 tags from the Propbank scheme (Bonial et al., 2010).

For Semantic Role Labeling (SRL) contexts, we try two types of extractions. One uses only the verb arguments as past contexts, another includes the verb, verb argument pairs. In all cases, the SRL output is concatenated into one string which is then separated by a start of sentence, end of sentence tokenizer pair. Results of the SRL extraction can be seen in table 6. We find the best performing model, of all models tested, is a model that uses the verb arguments of the three past contexts as context for the current dialogue. The performance is either better or on par with the baseline model regardless of which type of ROUGE measure and whether one considers recall or precision. Better precision, at the sake of recall, can be attained through SRL verb arguments of the previous 5 contexts. This strongly suggests a benefit of past contexts and that pre-processing the information of past contexts can be useful in increasing model performance.

5.5 Auto-summarization

In all of our experiments, we use human generated summaries as context. However, the transformer architecture trained with no past context information returns summaries of the last 100 seconds. Instead of requiring data collected via human-in-the-loop, we could instead use these automatically generated summaries as context for the model. Table 7 shows performance of 4 model variants trained either using human summaries or those automatically generated from the transformer architecture trained without previous summaries. In terms of recall, the human summaries result in better performance suggesting that a human-in-the-loop approach may result in better overall temporal summaries.

model	context	ROUGE-1	ROUGE-2	ROUGE-L
baseline	0	47.61/34.14	15.28/11.21	29.07/20.36
baseline$_{role}$	5	48.29/33.67	15.66/10.85	29.52/19.88
SRL	1	47.87/34.77	14.45/10.63	29.18/20.66
	3	**49.38**/33.80	**16.85**/11.41	**30.93**/20.40
	5	44.01/**36.77**	14.56/**12.40**	27.60/**22.56**
	10	47.40/34.06	15.34/11.25	29.10/20.44
SRLverb	1	46.49/36.27	13.66/10.62	28.60/21.64
	3	43.89/**38.88**	14.56/**12.95**	26.96/**23.48**
	5	44.90/**35.41**	13.69/10.93	26.98/20.80
	10	46.79/**35.98**	15.81/**12.14**	28.51/**21.45**
SRL$_{role}$	1	44.08/**38.32**	14.91/13.07	27.99/**23.74**
	3	44.18/**36.73**	14.36/11.96	27.47/**22.38**
	5	47.98/34.41	15.05/10.85	29.93/20.87
	10	47.64/**36.43**	15.66/**12.14**	28.82/**21.42**
SRLverb$_{role}$	1	46.47/34.74	14.37/10.91	28.96/**21.10**
	3	47.25/33.70	15.56/11.04	28.80/19.89
	5	46.20/**35.06**	15.26/11.54	28.80/**21.36**
	10	46.73/**36.01**	15.04/11.67	28.57/**21.33**

Table 6: R1, R2, and RL scores (recall/precision) for models that are trained with past contexts from semantic role labeling including verb object pair (SRLverb), with SRL objects (SRL) only.

summaries	context	ROUGE-1	ROUGE-2	ROUGE-L
human	5	46.81/34.55	13.87/10.13	28.50/20.50
auto	5	44.59/35.70	13.89/11.02	27.05/21.06
human$_{role}$	5	48.29/33.67	15.66/10.85	29.52/19.88
auto$_{role}$	5	46.67/36.50	14.07/11.18	28.66/21.95

Table 7: R1, R2, and RL scores (recall/precision) comparing human vs transformer generated summaries.

6 Discussion & Future work

In this work we present an analysis of the role of past context on summarizing 100 seconds of temporal meeting dialogue. We explore, in depth, the way in which past summaries can be used by a summarization model to generate abstractive summaries. Our work strongly suggests that context impacts model performance. We also find the way in which we represent previous summaries can impact metrics related to the quality of the abstractive summaries. We show that in certain conditions human generated summaries can improve over models with no contextual information. We then show that extracting meaningful content from past summaries can further boost model performance. Specifically, we found the verb arguments of a semantic role labeler provides the most performance improvement over our baseline models. We believe that this result provides a new direction for temporal summarization by suggesting that contextual information preceding the specific dialogue may be informative for the model in generating summaries.

To further analyze the summaries generated by the models we compare the summaries to the extractive text that was provided as input. Table 8 shows the ROUGE (Recall/Precision) measures for this comparison. We can make several observations from this table. We see that adding role information when there is no context helps improve the recall and precision (b,c in Table 8). We also observe that the human abstractive summaries (a) shows lowest recall and precision when compared to the extractive input text than those achieved via our temporal summarization models. This indicates that humans are generating summaries using tokens not present in the input which presents unique challenge to the summarization models. Another important observation we can make is that the precision of these models is high, suggesting that words in the model's abstract summary appear in the input. Recall, as expected, is low as many of the words in the input do not appear in the summary. We can also observe that adding more context information influences the SRL-based models in achieving better R2 & RL recall compared to the baseline.

model	context	ROUGE-1	ROUGE-2	ROUGE-L
(a) humans		18.73/49.62	5.09/12.91	10.84/28.65
(b) baseline	0	31.18/55.67	15.37/26.82	20.23/34.63
(c) baseline$_{role}$	0	**31.88**/58.16	15.87/28.01	20.38/36.14
(d) Keyword	1	29.79/**65.13**	15.57/33.15	19.30/40.79
	10	27.55/55.47	11.89/23.59	17.65/34.54
(e) Keyword$_{role}$	1	29.50/61.41	14.59/29.71	18.02/36.39
	10	28.33/64.43	14.80/**33.34**	18.89/**41.70**
(f) SRL	1	28.24/54.27	11.73/21.69	17.04/31.95
	10	31.01/59.66	16.07/30.70	19.94/37.29
(g) SRLverb	1	26.51/54.12	10.55/20.67	16.42/32.55
	10	29.25/58.46	15.28/29.84	18.94/36.48
(h) SRL$_{role}$	1	26.91/60.47	14.13/31.16	18.30/39.52
	10	31.69/61.88	**16.90**/32.94	**20.66**/39.37
(i) SRLverb$_{role}$	1	27.79/54.39	11.82/21.66	17.69/33.00
	10	30.74/60.84	14.67/28.74	18.91/36.16

Table 8: R1, R2, and RL scores (recall/precision) comparing model summaries to the extractive text of the meeting transcripts with context of 1 & 10.

There are limitations and clear future directions of this work. First, the model architecture we explored here is the standard BART summarization architecture. More recent models have achieved impressive performance on meeting summarizations (Feng et al., 2020; Zhu et al., 2020; Fabbri et al., 2021b). Exploring these architectures and adapting them for ITS scenario remains a promising avenue for the future work. This work also suggests that an architecture specifically aimed to capitalize on past summary information may be a promising line for our future work. When inspecting model performance, specifically when the role labels were not present, we found that the model tended to over-attend to previous contextual information. This may be mitigated by building an architecture that keeps dialogue and context information separate.

Our work provides a rather simplistic HITL (Human in the loop) approach for summarization. In this work, we integrate the summaries from the past as input to the models. While, the approach is simple, we have demonstrated that such a method of integrating context information could help improve the performance of the summarizer. Integrating human inputs into the inference pipeline is an interesting area for future work. Eventually, this system should be able to integrate human information seamlessly, requiring more experiments and analysis to understand how individuals are generating temporal summaries and how the model makes use of the past context for prediction.

One of the challenges is evaluating the quality of summaries in a scalable and automatic fashion. The ROUGE metrics are widely adopted for the purposes of summary evaluation (Lin, 2004). While numerous automated evaluation metrics exist for measuring how closely the generated summary matches with the ground-truth (Fabbri et al., 2021a) a metric for ITS scenario needs further research. Human evaluations are commonly adopted for measuring the summary quality. However, such an approach can be expensive and could also prove to be noisy when deployed over crowdsourcing environment. Recently Shapira et al. (2021) have highlighted the issue and provided an interactive evaluation of multi-document summaries. We intend to explore other types of evaluations and human judgements on ITS datasets in the future.

Incremental Temporal summarization is an emerging area of research and thus limited by data. We base all our analysis on the AMI-ITS dataset (Manuvinakurike et al., 2021). One aspect of this dataset is that summaries are generated by individuals who are seeing the 3 previous summaries generated by other crowdsource workers. These workers may be influenced by these previous summaries when generating their summaries of the last 100 seconds. Because of this, the summaries themselves may contain information about previous context making the addition of other contexts redundant and altering the extendability of these results. In the future, we intend to analyse and better understand how transformer models use previous context as well as how individuals determine what aspects of a meeting are important for incremental summarization.

References

Javed Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. 2015. TREC 2014 temporal summarization track overview. Technical report, NIST, Gaithersburg, MD.

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtić, Mark Hepple, and Robert Gaizauskas. 2016. The sensei annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 42–52.

Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. 2010. Propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39.

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.

Yun-Nung Chen and Florian Metze. 2012. Integrating intra-speaker topic modeling and temporal-based inter-speaker topic modeling in random walk for improved multi-party meeting summarization. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *TAC*.

Franck Dernoncourt, Mohammad Ghassemi, and Walter Chang. 2018. A repository of corpora for summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021b. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. *arXiv preprint arXiv:2106.00829*.

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *Dialogue*, 1:U2.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software*, pages 1–6.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

KM Hermann, T Kočiskỳ, E Grefenstette, L Espeholt, W Kay, M Suleyman, and P Blunsom. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, 28.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.*, volume 1.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3315–3326.

Ronilda C Lacson, Regina Barzilay, and William J Long. 2006. Automatic analysis of medical dialogue in the home hemodialysis domain: structure induction and summarization. *Journal of biomedical informatics*, 39(5):541–555.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019b. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821.

Ramesh Manuvinakurike, Saurav Sahay, Wenda Chen, and Lama Nachman. 2021. Incremental temporal summarization in multiparty meetings. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2014. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 301–310.

Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440.

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

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Massimo Poesio and Hannes Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue & Discourse*, 1(1).

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

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

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2021. Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 657–677, Online.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Diana Trandabat. 2011. Using semantic roles to improve summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 164–169.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Su Yan and Xiaojun Wan. 2014. Srrank: leveraging semantic roles for extractive multi-document summarization. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(12):2048–2058.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 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.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 194–203.