


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

Nicole M. Beckage, Shachi Kumar, Saurav Sahay, and Ramesh Manuvinakurike

Intel Labs, USA



Diachronic Language models (8LP)
Haofang Fan
Scientific Computing
haofang.fan@stud.uni-heidelberg.de

Contents



- Background
- Related work
- Dataset
- Model structure
- Models and results
- Conclusion

Background

Motivation

- The prevalence and needs of multi-party meetings
- Exploring human-in-the-loop approaches
- Use of advanced techniques

Goal

Investigate the ability to incrementally summarise meetings

Existing
problems

- The information evolves over time.
- How the summarisation tool make use of past summaries?
- How much context or past summaries should be used to generate a summary of the current dialogue?

Related Work

- News article: Structured content, clear thematic segments, factual and objective presentation.
 - News summarisation: Temporal summarisation in the context of summarising news article(Dang, 2008).
- Meeting dialogues: Dynamic, interactive, unstructured, involving multiple speakers and conversational elements.
 - Meeting summarisation: More complex, limited labeled training data.
 - Improve the accuracy of abstractive text summarisation(See, 2017).
 - Introduces a approach to process information from both audio and video components(Li, 2019).
- Incremental summarisation: Focuses on updating summaries of a text (or set of texts) as new information becomes available.
 - Propose a novel abstractive summary network HMNeT that adapts to the meeting scenario(Zhu, 2020).
 - Achieved SOTA performance on the AMI meeting corpus but has not been validated on incremental summarisation tasks.
- Incremental temporal summarisation: More specific on summarising events or topics that evolve over time.

Dataset



AMI(Carletta et al.,2005):

- Consist of conversations between 4 role-playing participant(PM, ID, UI and ME)
- Consisting of 100 hours of meeting recordings
- Each group of participants meet 4 times
- Each conversation forward from the previous sessions but on a new agenda
- Consist of extractive and abstractive summaries for the full conversation annotated by experts

AMI-ITS(Manuvinakurike,2021):

- Provide summaries for 100-second segments of the meetings
- Participants can see up to 3 summaries that captured the 3 preceding dialogue chunks
- Participants decide whether choose a dialogue or not to provide summary as down-stream materials

Dataset

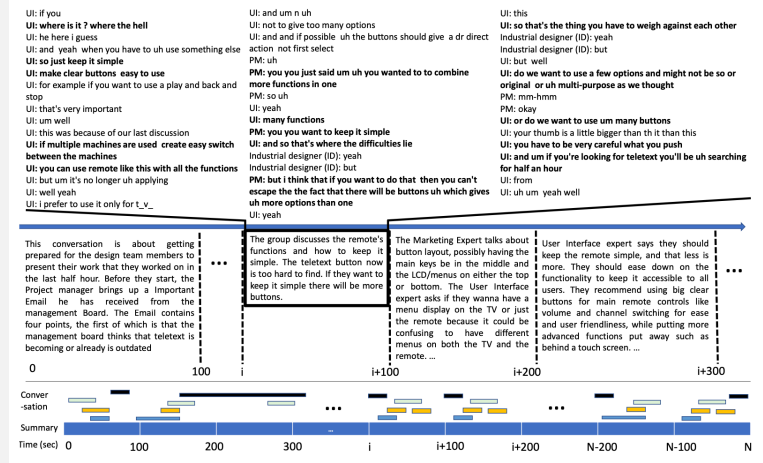


Figure1 (Beckage, 2021)

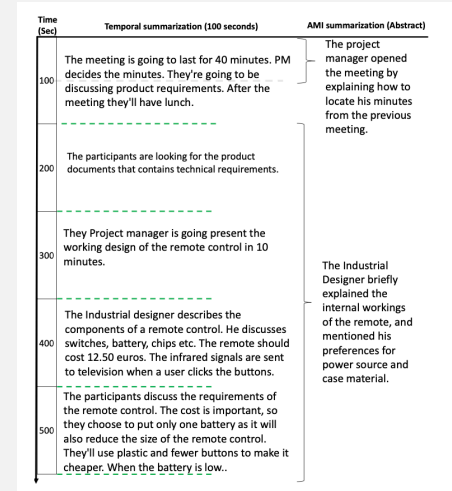


Figure2 (Manuvinakurike, 2021)

Model Structure

Primary:

- Develop an abstractive summarisation module for incremental temporal scenarios
- Explore the best methods for incorporating past summaries into the summarisation process
- Investigate the extent and effectiveness of using previous temporal summaries as context for accurate summarisation
- Draw conclusions about the influence of human summarisation practices on the performance of these models

Model input:

- Extractive meeting dialogues from AMI-ITS datasets (including human-judged sentences)
- Role information (role, e.g. 'Project Manager (PM):')

Model output:

- Summary of dialogue

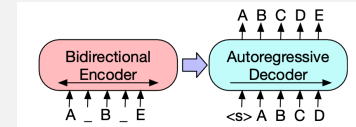
Context use:

Assume past summaries(human-generated) as the context, ranging from none to several.

Model Structure

BART as baseline model(Lewis,2020):

- Sequence-to-sequence transformer architecture
- Combine the bidirectional encoder of BERT with the left-to-right decoder of GPT
- Pre-train approach allows model to learn broad language features,
fine-tune approach makes model to adapt specific tasks
- Used for various natural language processing tasks



BART-large-CNN model(Wolf,2019) as pre-trained model:

- A variant of the BART model specifically optimised for summarisation tasks, particularly on the CNN/Daily Mail dataset

ROUGE

Recall: Indicating coverage

Precision: Indicating relevance

Models and Results

Fine-tune to dialogue

Aim: Investigate fine-tuning pre-trained model could improve the ability of incremental temporal summarisation.

Comparison among:

- BART-large-CNN model
- BART-large-CNN model with role information
- BART models fine-tuned on the AMI dataset
- BART models fine-tuned on the AMI dataset with role information

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-ITS 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.

Models and Results

Summaries and extractive texts

Aim: Investigate whether model performance changes when we truncate the input.

Comparison among:

- Models maintain the extractive text
- Models maintain the extractive text with role information
- Models maintain the summaries
- Models maintain the summaries with role information

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 where 10 contexts are used.

Models and Results

The effects of previous summaries

Aim: Research what extent previous (human) generated summaries improve the quality of the summaries.

Comparison among:

- Models with a various number of previous temporal summaries
- Models with a various number of previous temporal summaries with role information

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.

Models and Results

Capturing contexts

Aim: Investigate ways to capture on the relevant information from the past summaries.

Keyphrase:

Definition of context: 10 most important words or phrases from past summaries

How to extract keyphrases? — — Key-Bert model(Grootendorst,2020)

How to reduce redundancy and increase diversity? — — Maximal margin relevance(Carbonell and Goldstein,1998)

Model comparison:

- Baseline model
- Baseline model with role information as input
- Models trained with keyphrases
- Models trained with keyphrases and role information

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
	1	46.54/ 37.10	14.96/ 12.07	28.01/ 21.74
Keyphrase _{role}	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/precision) 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.

Models and Results

Capturing contexts

Aim: Investigate ways to capture on the relevant information from the past summaries.

Semantic role labeling:

Two types of extraction:

1. Use verb arguments as past context
2. Use verb, verb argument pairs

Model comparison:

- Baseline model (with and without role information)
- Model with SRL objects (with and without role information)
- Model with SRL including verb object pair (with and without role information)

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
SRL _{verb}	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
SRL _{verbrole}	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 (SRL_{verb}), with SRL objects (SRL) only.

Models and Results

Auto-summarisation

Aim: Investigate whether auto-summarisation can be used as context.

Comparison among:

- Model with human-generated summary(with or without role label)
- Model with auto-summarisation (with or without role label)

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.

Conclusion and Discussion

Conclusion:

- Context is found to significantly affect model performance
- Previous summaries can impact metrics related to the quality of the abstractive summaries
- Human-generated summaries can improve over models with no contextual information
- Verb arguments of a semantic role labeller provides the most performance improvement over the baseline model

Further research: Contextual information preceding specific dialogues could be informative for generating summaries

Validation:

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) SRL _{verb}	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) SRL _{verb} _{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.

Conclusion and Discussion

Contribution:

- Examines the influence of past context on the summarisation of 100-second segments of meeting dialogue.
- Contributes significantly to the study of incremental temporal summarisation.

Cons:

- Evaluation metrics for ITS scenarios are needed.
- Limitations of AMI-ITS dataset.
- Future work: New architecture in ITS scenarios, a structure that separates dialogue and contextual information, better integrate human information, other types of evaluations and human judgements on ITS datasets...

Questions:

- In the experiments of this paper, why is it that sometimes adding *role information* improves the performance of the model and sometimes it doesn't?
- When performing incremental temporal summarisation, what else could be considered to added in the model besides context and role information?

Thank you!

