Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization

Xiachong Feng[1], Xiaocheng Feng[1,2], Libo Qin[1], Bing Qin[1,2], Ting Liu[1,2]

1 Harbin Institute of Technology, 2 Peng Cheng Laboratory

Dialogue Summarization

- Dialogue summarization aims to generate a succinct summary while retaining essential information of the dialogue.

Dialogue

Blair:	Remember we are seeing the wedding planner after work
Chuck:	Sure, where are we meeting her?
Blair:	At Nonna Rita's
Chuck:	I want to order seafood tagliatelle
Blair:	Haha why not
Chuck:	We remmber spaghetti pomodoro disaster from our last meeting
Blair:	Omg it was over her white blouse
Chuck:	:D
Blair:	:P

Summary

Blair and Chuck are going to meet the wedding planner after work at Nonna Rita's. The tagliatelle served at Nonna Rita's are very good.

A Good Summary?

Peyrard (2019): a good summary is intuitively related to three aspects

Informativeness | **Redundancy** | **Relevance**

Dialogue	
Blair:	Remember we are seeing the **wedding planner** after work
Chuck:	Sure, where are we meeting her?
Blair:	At **Nonna Rita's**
Chuck:	I want to order **seafood tagliatelle**
Blair:	Haha why not
Chuck:	We remmber **spaghetti pomodoro disaster** from our last meeting
Blair:	Omg it was over her white blouse
Chuck:	I'll make time for it
Blair:	Great!

(a) Keywords Extraction

Dialogue	
Blair:	Remember we are seeing the wedding planner after work
Chuck:	Sure, where are we meeting her?
Blair:	At Nonna Rita's
Chuck:	I want to order seafood tagliatelle
Blair:	***Haha why not***
Chuck:	We remmber spaghetti pomodoro disaster from our last meeting
Blair:	Omg it was over her white blouse
Chuck:	***I'll make time for it***
Blair:	***Great!***

(b) Redundancy Detection

Dialogue		
Blair:	Remember we are seeing the wedding planner after work	
Chuck:	Sure, where are we meeting her?	
Blair:	At Nonna Rita's	[Topic 1]
Chuck:	I want to order seafood tagliatelle	
Blair:	Haha why not	
Chuck:	We remmber spaghetti pomodoro disaster from our last meeting	[Topic 2]
Blair:	Omg it was over her white blouse	
Chuck:	I'll make time for it	[Topic 3]
Blair:	Great!	

(c) Topic Segmentation

Summary
Blair and Chuck are going to meet the **wedding planner** after work at **Nonna Rita's**. The **tagliatelle** served at **Nonna Rita's** are very good.
$\underbrace{\text{Blair and Chuck are going to meet the wedding planner after work at Nonna Rita's.}}_{\text{[Topic 1]}}$ $\underbrace{\text{The tagliatelle served at Nonna Rita's are very good.}}_{\text{[Topic 2]}}$

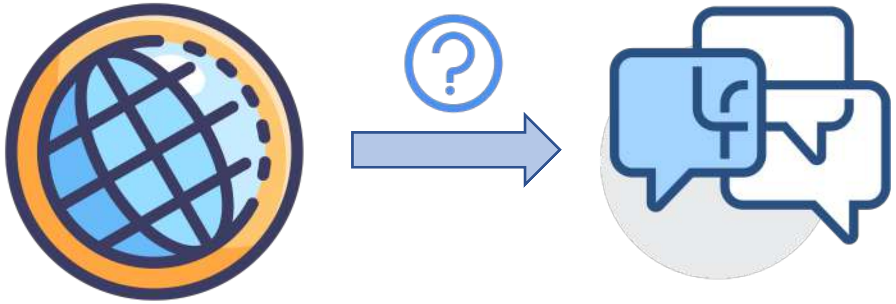
Related Works

- **For informativeness**
  - Linguistically specific words
  - Domain terminologies
  - Topic words
- **For redundancy**
  - Similarity-based methods to annotate redundant utterances
- **For relevance**
  - Topic segmentation

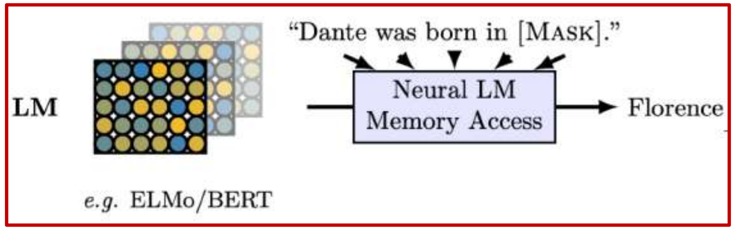
Problems

- **Relied on human annotations.**
  - labor-consuming

- **Obtained via open-domain toolkits**
  - Dialogue agnostic
  - not suitable for dialogues



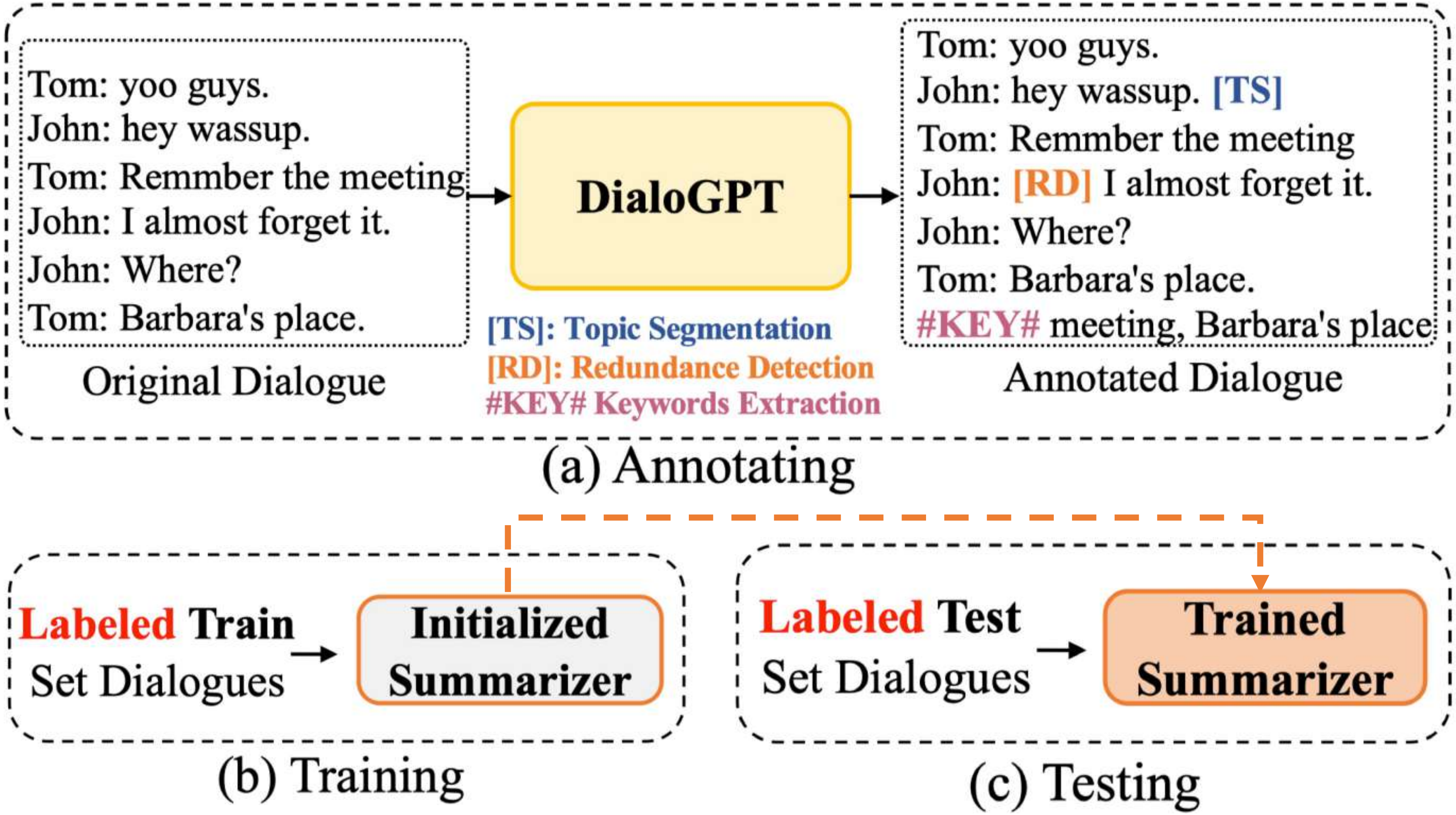
Pre-trained Language Models



Knowledge Base Prompt Tuning Zero-shot learning

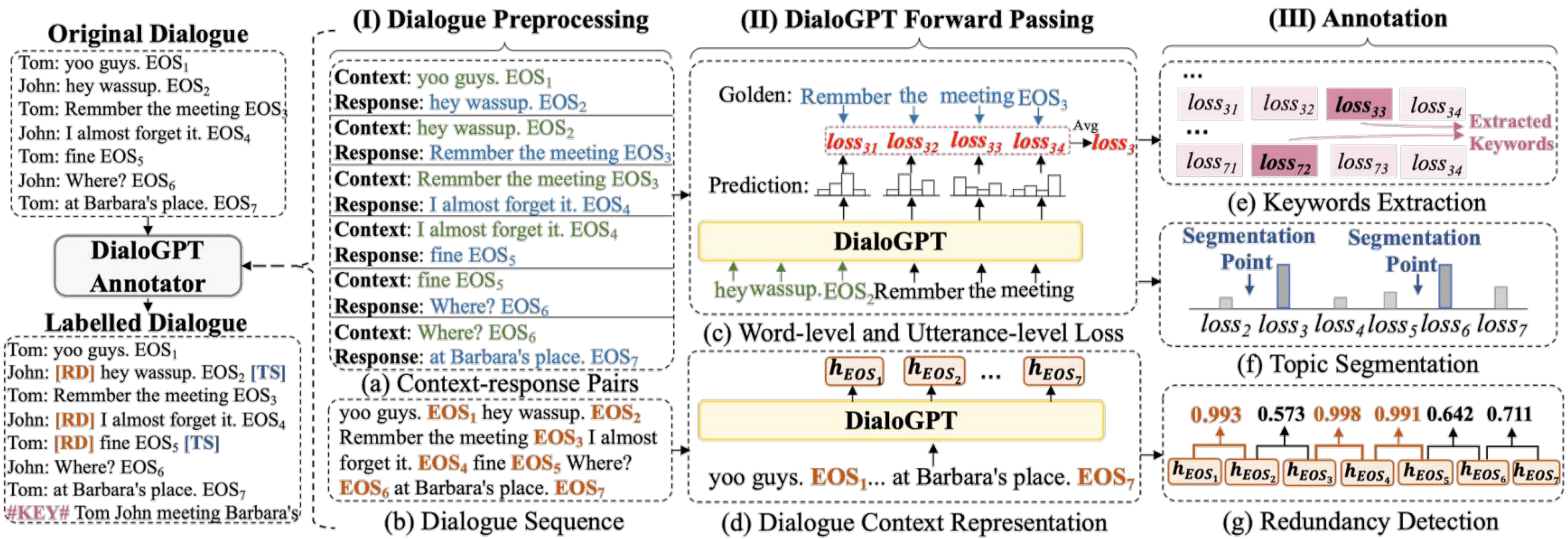
Pre-encoded Knowledge

DialoGPT Annotator



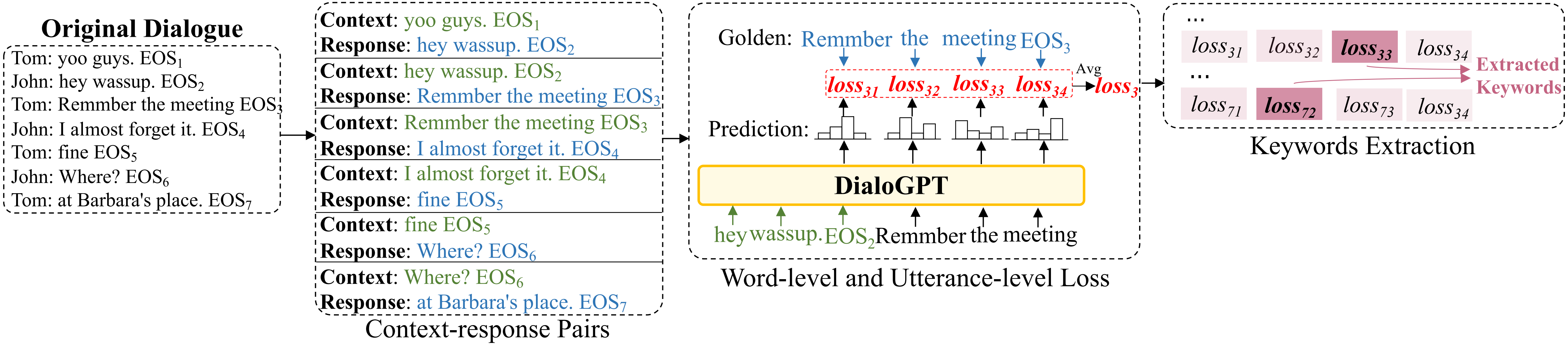
Overview

- **Keywords Extraction:** Extracts unpredictable words as keywords.
- **Topic Segmentation:** Inserts a topic segmentation point before one utterance if it is unpredictable.
- **Redundancy Detection:** Detects utterances that are useless for context representation as redundant.



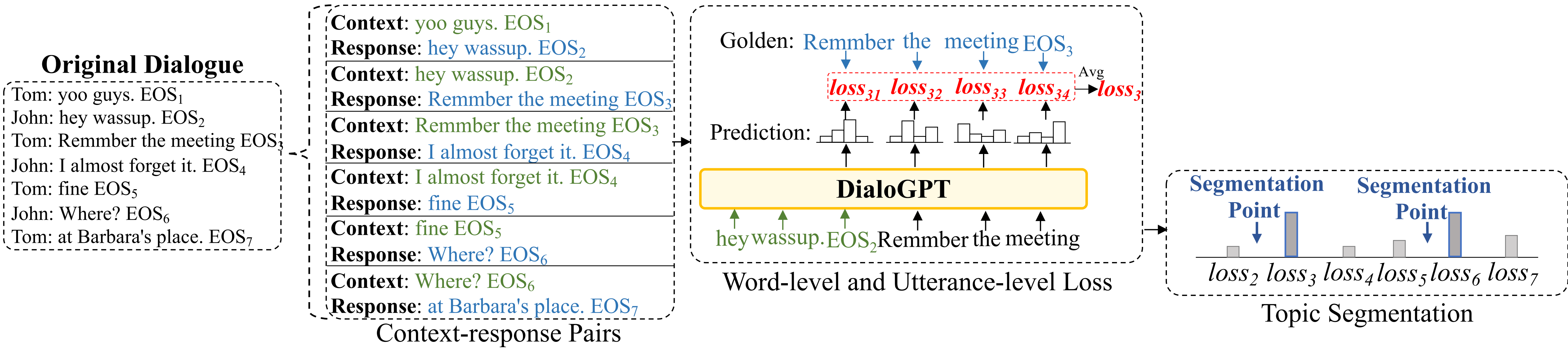
Keywords Extraction: DialoGPT$_{KE}$

- **Motivation:** if one word in the golden response is difficult to be inferred from DialoGPT, we assume that it contains high information and can be viewed as a keyword.
- Extracts unpredictable words as keywords.



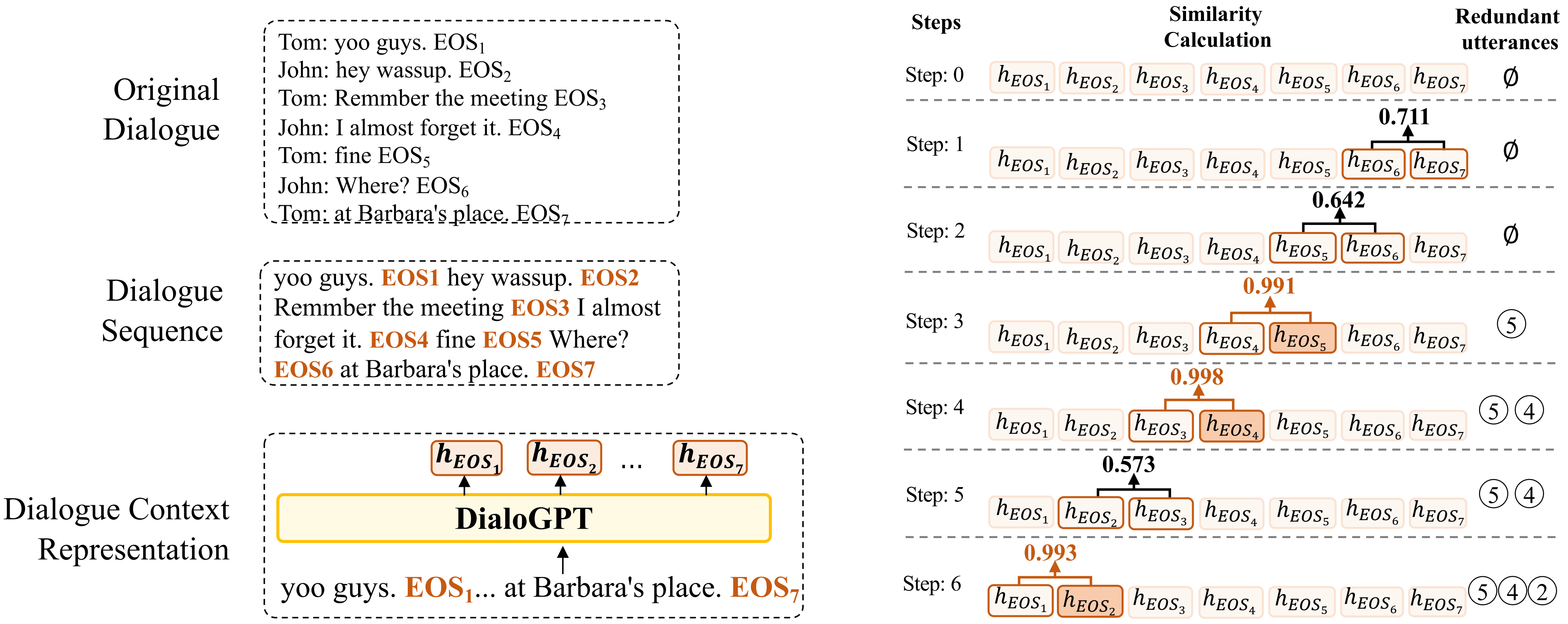
Topic Segmentation: DialoGPT$_{TS}$

- **Motivation:** if the response is difficult to be predicted given the context based on DialoGPT, we assume the response may belong to another topic and there is a topic segmentation between the context and response.
- Inserts a topic segmentation point before one utterance if it is unpredictable.

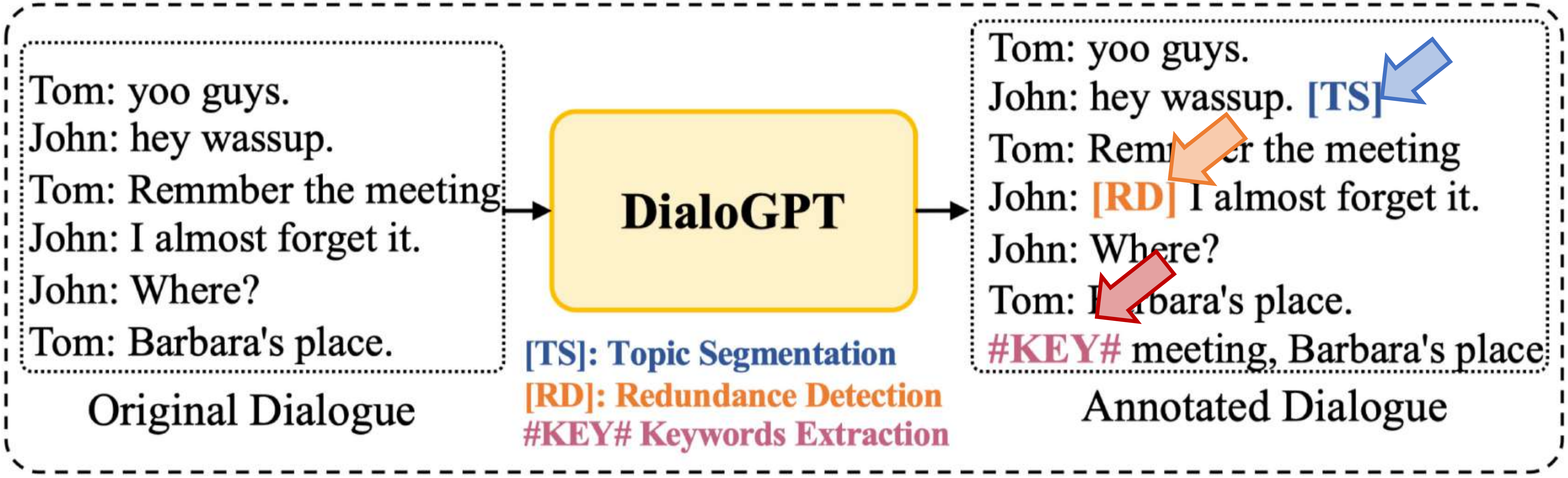


Redundancy Detection: DialoGPT$_{RD}$

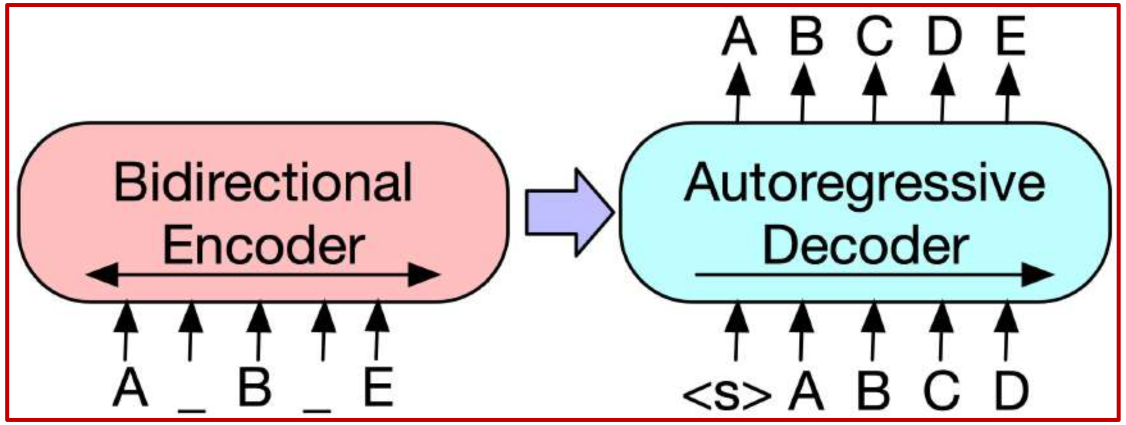
- **Motivation:** If one utterance brings brings little information and has small effects on predicting the response, this utterance becomes a redundant utterance.
- Detects utterances that are useless for context representation as redundant.



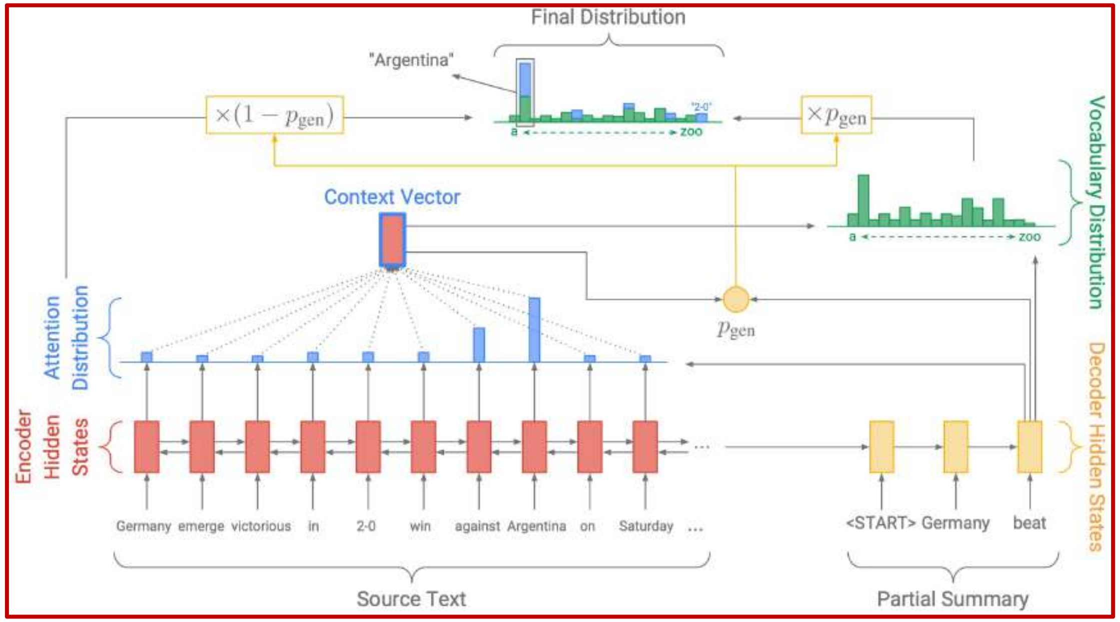
Annotation Tags



Summarizer



BART
Pre-trained



PGN
Non pre-trained

Dataset and Metrics

- **Datasets**
  - SAMSum
  - AMI

		Train	Valid	Test
SAMSum	#	14732	818	819
	Avg.Turns	11.13	10.72	11.24
	Avg.Tokens	120.26	117.46	122.71
	Avg.Sum	22.81	22.80	22.47
AMI	#	97	20	20
	Avg.Turns	310.23	345.70	324.40
	Avg.Tokens	4859.52	5056.25	5257.80
	Avg.Sum	323.74	321.25	328.20

Statistics for SAMSum and AMI datasets

- **Evaluation Metrics**
  - ROUGE
  - BERTScore

Automatic Evaluation

Model	R-1	R-2	R-L
Extractive			
LONGEST-3	32.46	10.27	29.92
TextRank	29.27	8.02	28.78
Abstractive			
Transformer	36.62	11.18	33.06
D-HGN	42.03	18.07	39.56
TGDGA	43.11	19.15	40.49
DialoGPT	39.77	16.58	38.42
MV-BART	53.42	27.98	**49.97**††
Ours			
BART	52.98	27.67	49.06
BART($\mathcal{D}_{\mathrm{KE}}$)	**53.43**††	**28.03**††	49.93
BART($\mathcal{D}_{\mathrm{RD}}$)	53.39	28.01	49.49
BART($\mathcal{D}_{\mathrm{TS}}$)	53.34	27.85	49.64
BART($\mathcal{D}_{\mathrm{ALL}}$)	**53.70**†	**28.79**†	**50.81**†

Test set results on the SAMSum dataset

Model	R-1	R-2	R-L
Extractive			
TextRank	35.19	6.13	15.70
SummaRunner	30.98	5.54	13.91
Abstractive			
UNS	37.86	7.84	13.72
TopicSeg	**51.53**††	12.23	**25.47**†
HMNet	**52.36**†	**18.63**†	24.00
Ours			
PGN	48.34	16.02	23.49
PGN($\mathcal{D}_{\mathrm{KE}}$)	50.22	17.74	24.11
PGN($\mathcal{D}_{\mathrm{RD}}$)	50.62	16.86	24.27
PGN($\mathcal{D}_{\mathrm{TS}}$)	48.59	16.07	24.05
PGN($\mathcal{D}_{\mathrm{ALL}}$)	50.91	**17.75**††	**24.59**††

Test set results on the AMI dataset

SAMSum		AMI	
Model	**BS**	**Model**	**BS**
BART	86.91	PGN	80.51
MV-BART	88.46	HMNet	82.24
BART($\mathcal{D}_{\mathrm{ALL}}$)	**90.04**	PGN($\mathcal{D}_{\mathrm{ALL}}$)	**82.76**

BERTScore

Human Evaluation

	Model	Info.	Conc.	Cov.
SAMSum	Golden	4.37	4.26	4.27
	BART	3.66	3.65	3.66
	MV-BART	3.85	3.76	3.88
	BART($\mathcal{D}_{KE}$)	3.88	3.77	3.79
	BART($\mathcal{D}_{RD}$)	3.74	**3.98** †	3.89
	BART($\mathcal{D}_{TS}$)	**3.95**††	3.76	**4.01**††
	BART($\mathcal{D}_{ALL}$)	**4.05**†	**3.78**††	**4.08**†
AMI	Golden	4.70	3.85	4.35
	PGN	2.92	3.08	2.70
	HMNet	**3.52**†	2.40	**3.40**†
	PGN($\mathcal{D}_{KE}$)	3.20	3.08	3.00
	PGN($\mathcal{D}_{RD}$)	3.15	**3.25**†	3.00
	PGN($\mathcal{D}_{TS}$)	3.05	**3.10**††	**3.17**††
	PGN($\mathcal{D}_{ALL}$)	**3.33**††	**3.25**†	3.10

model can perform better in coverage

model can get the best score in conciseness

Effect of DialoGPT$_{KE}$

- Entities play an important role in the summary generation.
- Combined with DialoGPT embeddings, KeyBERT can get better results.

Method	R-1	R-2	R-L
Rule-Based Methods			
Entities	53.36	27.71	49.69
Nouns and Verbs	52.75	27.48	48.82
Traditional Methods			
TextRank	53.29	27.66	49.33
Topic words	53.28	27.76	49.59
Pre-trained Language Model-Based Methods			
KeyBERT			
w/ BERT emb	52.39	27.14	48.52
w/ DialoGPT emb	53.14	27.25	49.42
Ours			
DialoGPT$_{KE}$	**53.43**	**28.03**	**49.93**

Intrinsic Evaluation For Keywords

- View reference summary words as golden keywords
- Both TextRank and Entities perform poorly in recall
- Our method can extract more diverse keywords.

Method	Precision	Recall	F_1
TextRank	47.74%	17.44%	23.22%
Entities	**60.42%**	17.80%	25.38%
DialoGPT$_{KE}$	33.20%	**29.49%**	**30.31%**

Effect of DialoGPT$_{RD}$

- Rule-based method: annotates utterances without noun, verb and adjective as redundant.
- Our method shows more advantages for long and verbose meeting transcripts in the AMI.

Model	R-1	R-2	R-L
SAMSum			
Rule-based	53.00	27.71	**49.68**
DialoGPT$_{RD}$	**53.39**	**28.01**	49.49
AMI			
Rule-based	50.19	16.45	23.95
DialoGPT$_{RD}$	**50.62**	**16.86**	**24.27**

Effect of DialoGPT$_{TS}$

- Our method can get comparable results with the strong baseline C99(w/ DialoGPT emb).

Model	R-1	R-2	R-L
SAMSum			
C99			
w/ BERT emb	52.80	27.78	49.50
w/ DialoGPT emb	53.33	**28.04**	49.39
DialoGPT$_{TS}$	**53.34**	27.85	**49.64**
AMI			
Golden	50.28	19.73	24.45
C99			
w/ BERT emb	48.53	15.84	23.63
w/ DialoGPT emb	**49.22**	**16.79**	23.88
DialoGPT$_{TS}$	48.59	16.07	**24.05**

Ablation Studies for Annotations

- For both datasets, training summarizers based on datasets with two of three annotations can surpass corresponding summarizers that are trained based on datasets with one type of annotation.
- Summarizers that are trained on D_{KE+TS} still get improvements on both datasets.

Model	R-1	R-2	R-L
Ours			
BART	52.98	27.67	49.06
BART($\mathcal{D}_{\text{KE}}$)	53.43	28.03	49.93
BART($\mathcal{D}_{\text{RD}}$)	53.39	28.01	49.49
BART($\mathcal{D}_{\text{TS}}$)	53.34	27.85	49.64
BART($\mathcal{D}_{\text{KE+RD}}$)	53.56	28.65	50.55
BART($\mathcal{D}_{\text{KE+TS}}$)	53.51	28.13	50.00
BART($\mathcal{D}_{\text{RD+TS}}$)	53.64	28.33	50.13
BART($\mathcal{D}_{\text{ALL}}$)	**53.70**	**28.79**	**50.81**

Model	R-1	R-2	R-L
Ours			
PGN	48.34	16.02	23.49
PGN($\mathcal{D}_{\text{KE}}$)	50.22	17.74	24.11
PGN($\mathcal{D}_{\text{RD}}$)	50.62	16.86	24.27
PGN($\mathcal{D}_{\text{TS}}$)	48.59	16.07	24.05
PGN($\mathcal{D}_{\text{KE+RD}}$)	50.74	17.11	24.52
PGN($\mathcal{D}_{\text{KE+TS}}$)	50.69	16.83	24.33
PGN($\mathcal{D}_{\text{RD+TS}}$)	50.70	16.96	24.38
PGN($\mathcal{D}_{\text{ALL}}$)	**50.91**	**17.75**	**24.59**

Conclusion

- We investigate to use DialoGPT as unsupervised annotators for dialogue summarization, including keywords extraction, redundancy detection and topic segmentation.
- Experimental results show that our method consistently obtains improvements upon pre-traind summarizer (BART) and non pre-trained summarizer (PGN) on both datasets.
- Combining all three annotations, our summarizer can achieve new state-of-the-art performance on the SAMSum dataset.

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

SCIR
Research Center for Social Computing and Information Retrieval
理解语言 认知社会

Thanks!