

Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization

Xiachong Feng¹, Xiaocheng Feng^{1,2}, Libo Qin¹, Bing Qin^{1,2}, Ting Liu^{1,2}
1 Harbin Institute of Technology 2 Peng Cheng Laboratory

Introduction

Dialogue Summarization

•Generate a succinct summary while retaining essential information of the dialogue.

Previous Works

• Previous models usually encode the dialogue with additional semantic features.

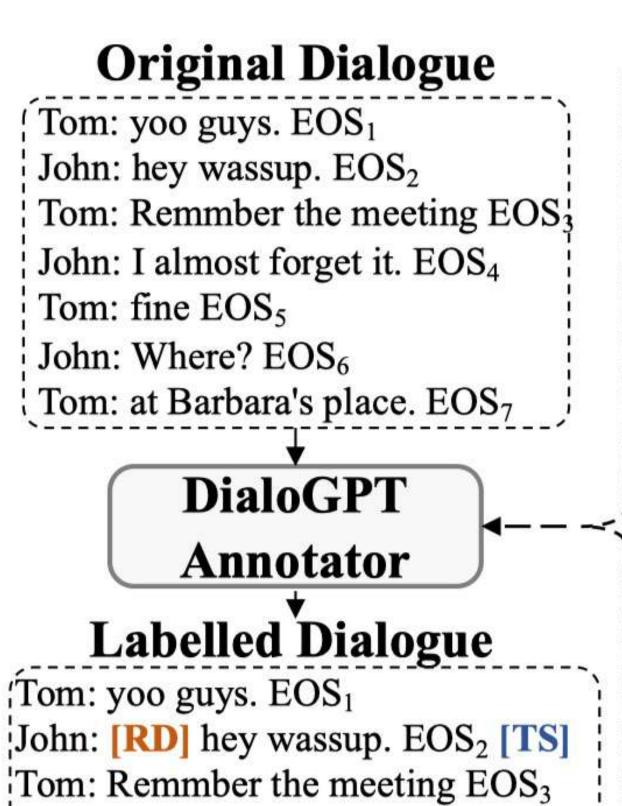
Problems

 Additional features are obtained via opendomain toolkits or relied on human annotations.

Our Solution

 View the pre-trained DialoGPT as an unsupervised dialogue annotator to label three features.

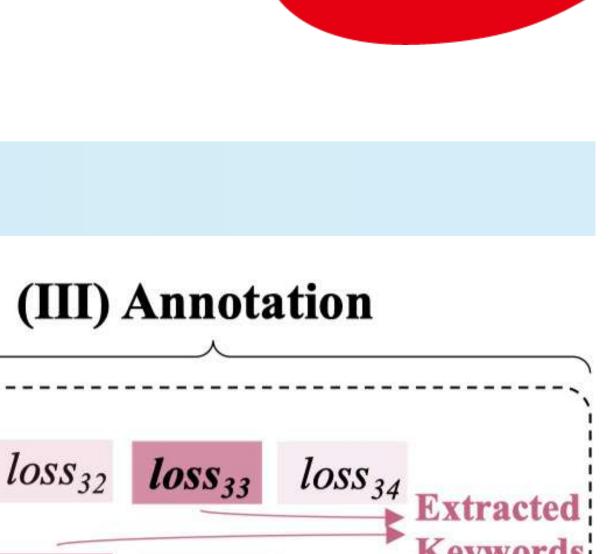
DialoGPT Annotator



Tom: yoo guys. EOS₁
John: [RD] hey wassup. EOS₂ [TS]
Tom: Remmber the meeting EOS₃
John: [RD] I almost forget it. EOS₄
Tom: [RD] fine EOS₅ [TS]
John: Where? EOS₆
Tom: at Barbara's place. EOS₇
#KEY# Tom John meeting Barbara's

(I) Dialogue Preprocessing Context: yoo guys. EOS₁ Response: hey wassup. EOS₂ Context: hey wassup. EOS₂ Response: Remmber the meeting EOS₃ Context: Remmber the meeting EOS₃ Response: I almost forget it. EOS₄ Context: I almost forget it. EOS₄ Response: fine EOS₅ Context: fine EOS₅ **Response: Where? EOS**₆ Context: Where? EOS₆ Response: at Barbara's place. EOS₇ (a) Context-response Pairs yoo guys. **EOS**₁ hey wassup. **EOS**₂ Remmber the meeting **EOS**₃ I almost forget it. **EOS**₄ fine **EOS**₅ Where? EOS₆ at Barbara's place. EOS₇

(II) DialoGPT Forward Passing



(e) Keywords Extraction

Segmentation Segmentation

Point Point Point I Point

(f) Topic Segmentation

0.993 0.573 0.998 0.991 0.642 0.711 h_{EOS_1} h_{EOS_2} h_{EOS_3} h_{EOS_4} h_{EOS_5} h_{EOS_6}

(g) Redundancy Detection

Experiments

Analysis

(b) Dialogue Sequence

Effect of DialoGPT_{KE}

Method

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			
Tradition	nal Meth	ods				
TextRank	53.29	27.66	49.33			
Topic words	53.28	27.76	49.59			
Pre-trained Languag	e Model	-Based N	1ethods			
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			

TextRank	47.74%	17.44%	23.22%
Entities	60.42%	17.80%	25.38%
DialoGPT _{KE}	33.20%	29.49%	30.31%
Effec	ct of Dia	loGPT _R	<u>D</u>
Model	R-1	R-2	R-L
	SAMSı	um	
Rule-based	53.00	27.71	49.68
DialoGPT _R	D 53.39	28.01	49.49
	AMI		
Rule-based	50.19	16.45	23.95
DialoGPT _R	D 50.62	16.86	24.27

Precision

Recall

Ablation Study

SAMSum Dialogue Dataset AMI Meeting Dataset

Model	R-1	R-2	R-L	Model	R-1	R-2	R-L	
	Ours			Ours				
BART	52.98	27.67	49.06	PGN	48.34	16.02	23.49	
$\mathrm{BART}(\mathcal{D}_{\mathrm{KE}})$	53.43	28.03	49.93	$PGN(\mathcal{D}_{KE})$	50.22	17.74	24.11	
$BART(\mathcal{D}_{RD})$	53.39	28.01	49.49	$\mathrm{PGN}(\mathcal{D}_{\mathrm{RD}})$	50.62	16.86	24.27	
$BART(\mathcal{D}_{Ts})$	53.34	27.85	49.64	$\mathrm{PGN}(\mathcal{D}_{\mathrm{Ts}})$	48.59	16.07	24.05	
$BART(\mathcal{D}_{KE+RD})$	53.56	28.65	50.55	$PGN(\mathcal{D}_{KE+RD})$	50.74	17.11	24.52	
$BART(\mathcal{D}_{KE+TS})$	53.51	28.13	50.00	$PGN(\mathcal{D}_{KE+TS})$	50.69	16.83	24.33	
$BART(\mathcal{D}_{RD+TS})$	53.64	28.33	50.13	$PGN(\mathcal{D}_{RD+Ts})$	50.70	16.96	24.38	
$\mathrm{BART}(\mathcal{D}_{\mathrm{ALL}})$	53.70	28.79	50.81	$\operatorname{PGN}(\mathcal{D}_{\operatorname{ALL}})$	50.91	17.75	24.59	

Conclusion

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

Automatic Evaluation

SAMSum Dialogue Dataset

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			
$BART(\mathcal{D}_{KE})$	53.43 ^{††}	28.03 ^{††}	49.93			
$BART(\mathcal{D}_{RD})$	53.39	28.01	49.49			
$BART(\mathcal{D}_{Ts})$	53.34	27.85	49.64			
$BART(\mathcal{D}_{AII})$	53.70 [†]	28.79 [†]	50.81 [†]			

AMI Meeting Dataset

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

Human Evaluation

	Model	Info.	Conc.	Cov.
	Golden	4.37	4.26	4.27
ı	BART	3.66	3.65	3.66
nn	MV-BART	3.85	3.76	3.88
MS	$BART(\mathcal{D}_{KE})$	3.88	3.77	3.79
SAMSum	$BART(\mathcal{D}_{RD})$	3.74	3.98 [†]	3.89
	$BART(\mathcal{D}_{Ts})$	$3.95^{\dagger\dagger}$	3.76	$4.01^{\dagger\dagger}$
	$\mathrm{BART}(\mathcal{D}_{\mathrm{ALL}})$	4.05^{\dagger}	$3.78^{\dagger\dagger}$	4.08^{\dagger}

	Model	Info.	Conc.	Cov.
	Golden	4.70	3.85	4.35
	PGN	-2.92^{-}	3.08	2.70^{-}
	HMNet	3.52^{\dagger}	2.40	3.40^{\dagger}
AMI	$PGN(\mathcal{D}_{KE})$	3.20	3.08	3.00
7	$PGN(\mathcal{D}_{RD})$	3.15	3.25^{\dagger}	3.00
	$PGN(\mathcal{D}_{Ts})$	3.05	$3.10^{\dagger\dagger}$	$3.17^{\dagger\dagger}$
	$PGN(\mathcal{D}_{ALL})$	3.33 ^{††}	3.25^{\dagger}	3.10

Effect of DialoGPT_{TS}

Model	R-1	R-2	R-L
SA	MSum		
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

Model	R-1	R-2	R-L
A	MI		
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