EACL 2023 Tutorial

Summarization of Dialogues and Conversations At Scale

Part 2: Pretraining and Models

Chenguang Zhu
Microsoft Cognitive Services Research

Challenges of Dialogue Summarization

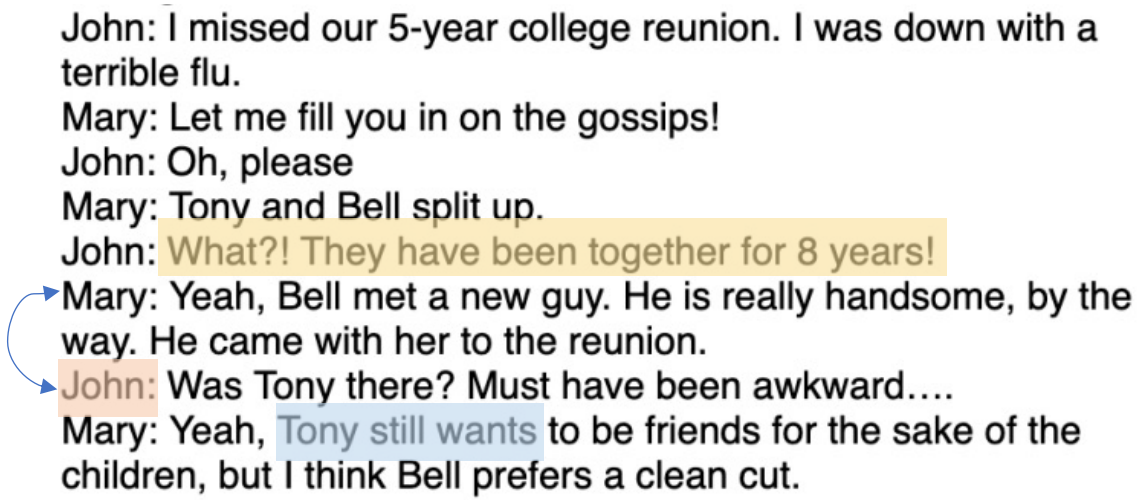
- Multiple participants
  - 2-20 participants are typical
  - Each participant has different semantic style, point of view, etc.
- Long conversation and long summary
  - The transcript of a one-hour meeting typically contains 5K-10K words
  - Summary can be 200-500 words depending on style
- Distinct Content
  - Domain-specific knowledge
  - Reference to participants
  - Colloquial style and error from speech recognition

Pretraining

- Lack of large-scale high-quality dialogue summarization datasets
  - Privacy issue: Many dialogues are personal or business related
- LM Pretraining is a powerful method to alleviate lack of downstream task data
  - Self-supervised learning such as masked language model (MLM), denoising auto-encoder (DAE)
  - Models: BERT, RoBERTa, T5, BART, etc.
- Existing pre-trained LM are primarily for documents, not proper for dialogues

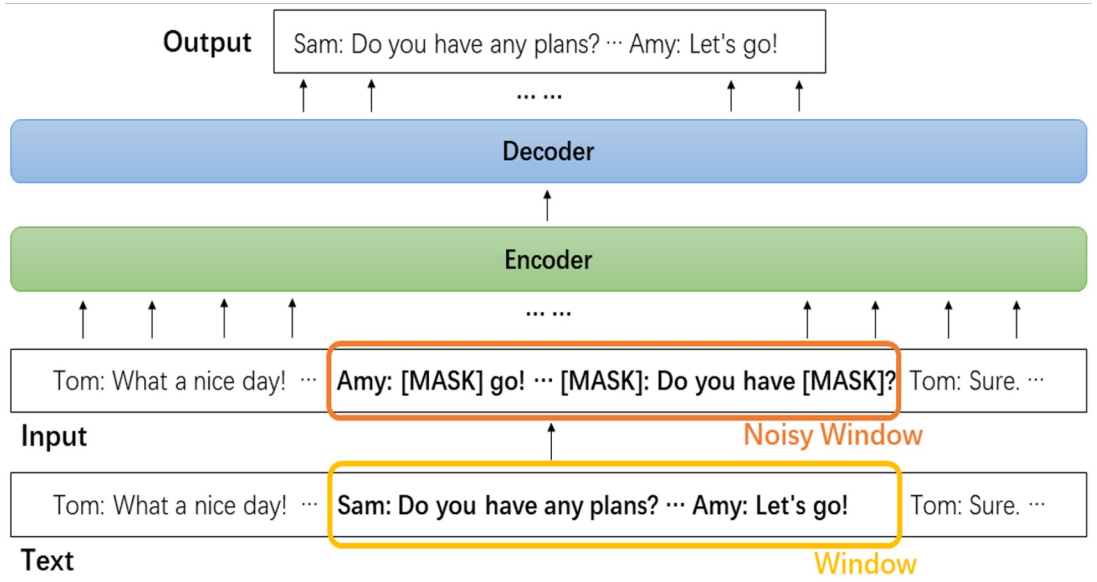
Designing Pre-training Tasks for Dialogues

- Modify masking and noising methods to focus on turn-based transcript structure, e.g.,
  - Mask the whole turn
  - Randomly shuffle neighboring turns
  - Mask speakers
- Goal: recover the original turn & speaker info



Pretraining: DialogLM

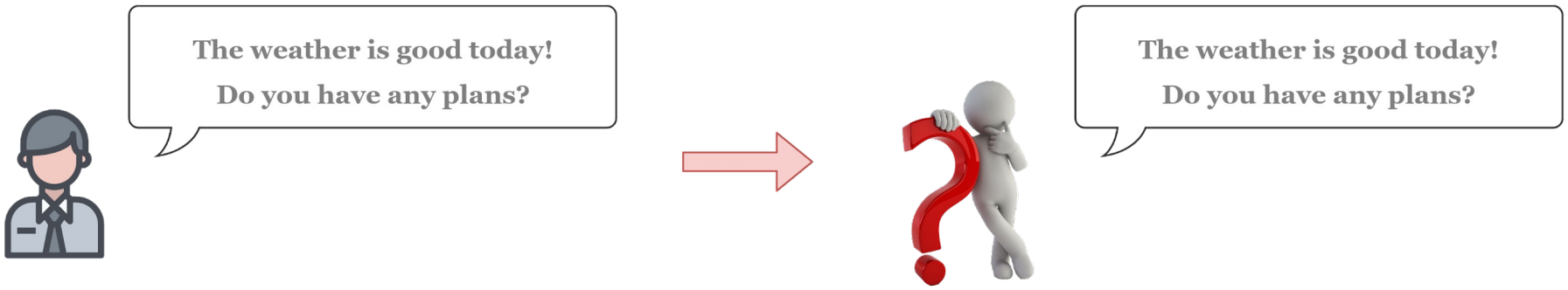
- Window-based Denoising
  - A window consists of consecutive turns
  - All noises are applied to this window
  - Input: the whole dialogue with the noisy window
  - Output: the denoised window



[1] *DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization*
M. Zhong et al. AAAI 2022

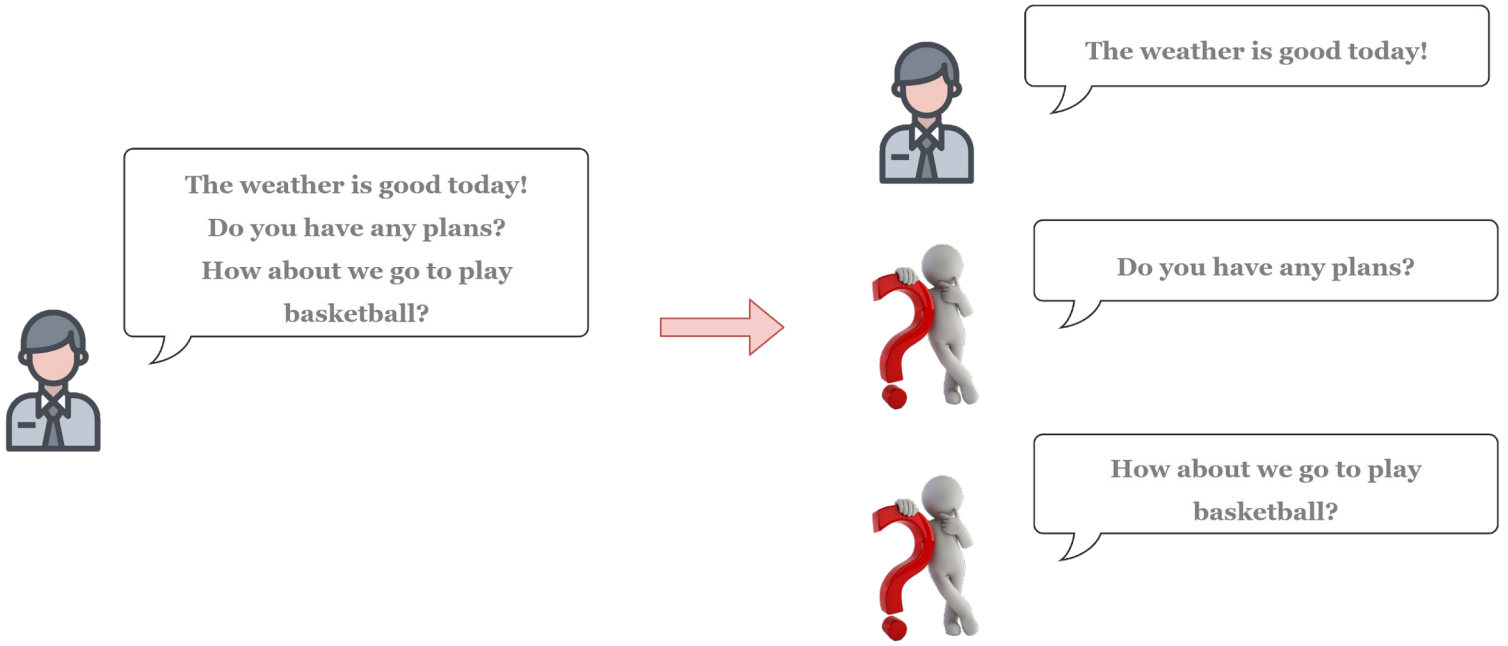
Pretraining: DialogLM

- Noise 1: Speaker Mask
- Goal: Help the model identify the speaker



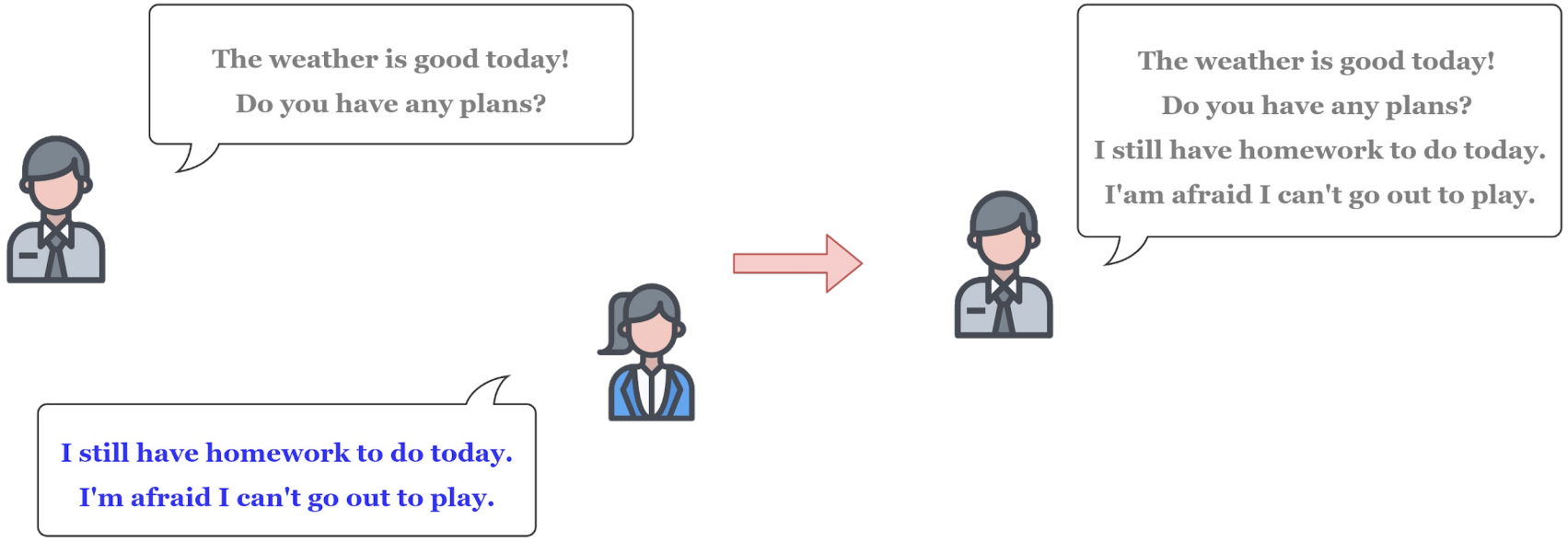
Pretraining: DialogLM

- Noise 2: Turn Splitting
- Goal: Help the model identify the speaker and the boundary between turns



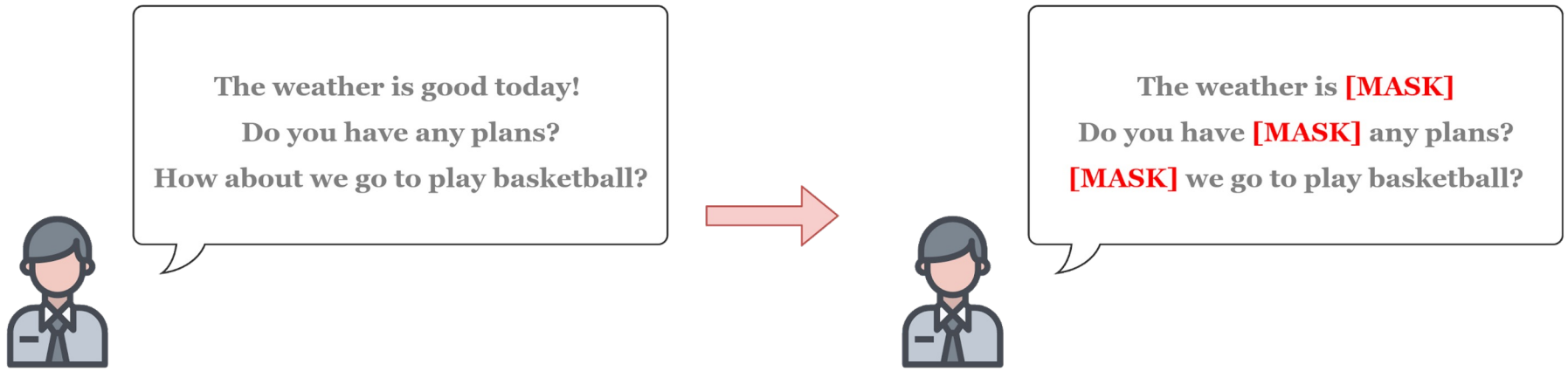
Pretraining: DialogLM

- Noise 3: Turn Merging
- Goal: Help the model identify the speaker and the boundary between turns



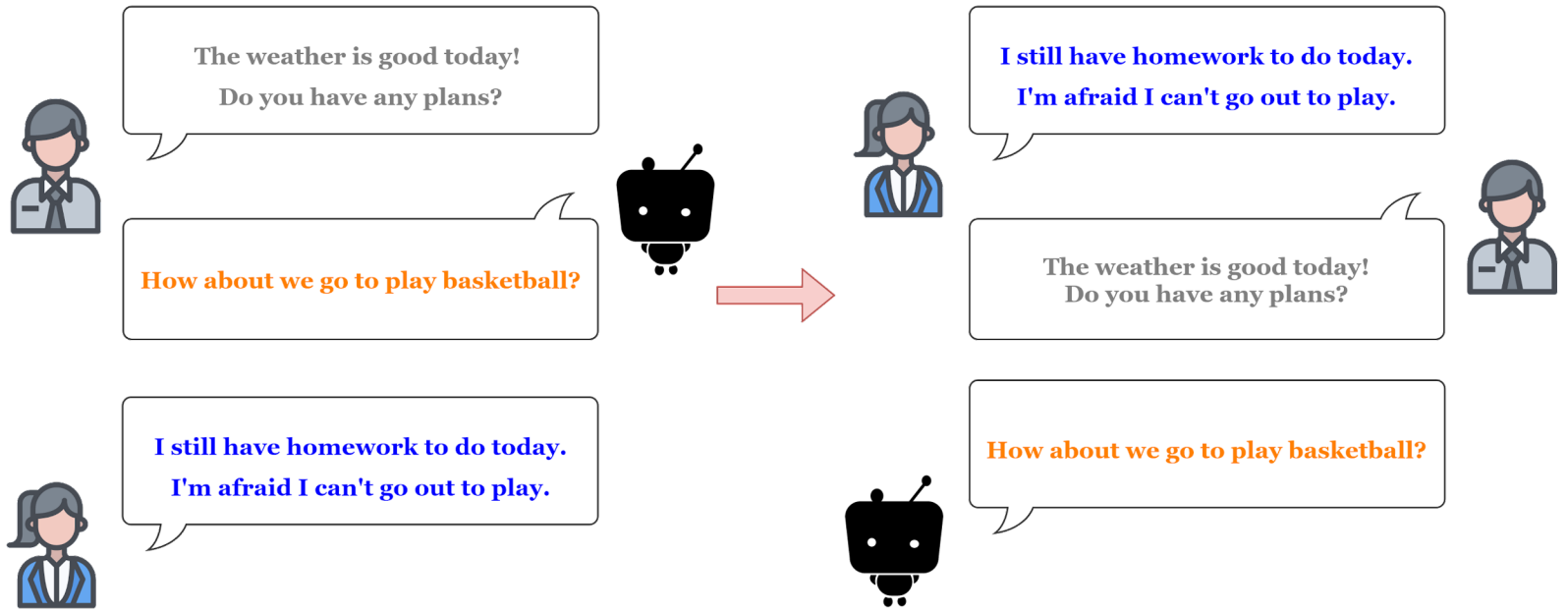
Pretraining: DialogLM

- Noise 4: Text Infilling
- Goal: Help the model understand the content of the utterance



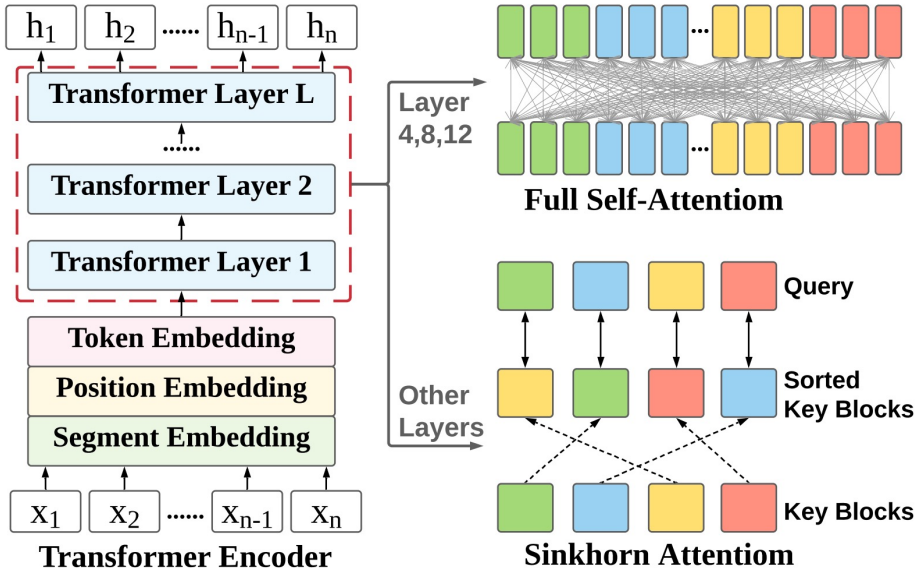
Pretraining: DialogLM

- Noise 5: Turn Permutation
- Goal: Help the model understand the order of turns in the dialogue



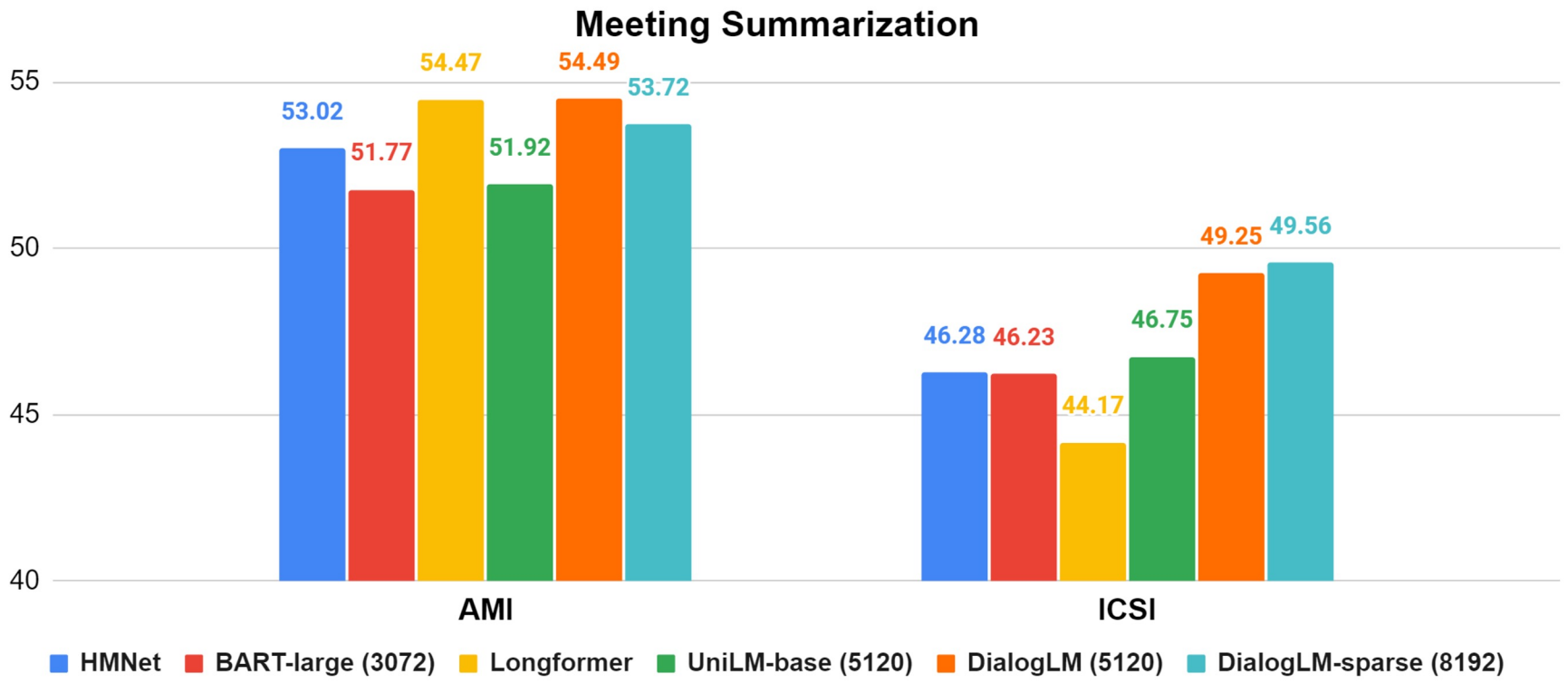
Pretraining: DialogLM

- Continue pretraining UniLM on MediaSum and OpenSubtitles datasets with ~600K dialogues
- Extend the input limit from 512 to 8,192 tokens by Sinkhorn sparse block-based attention



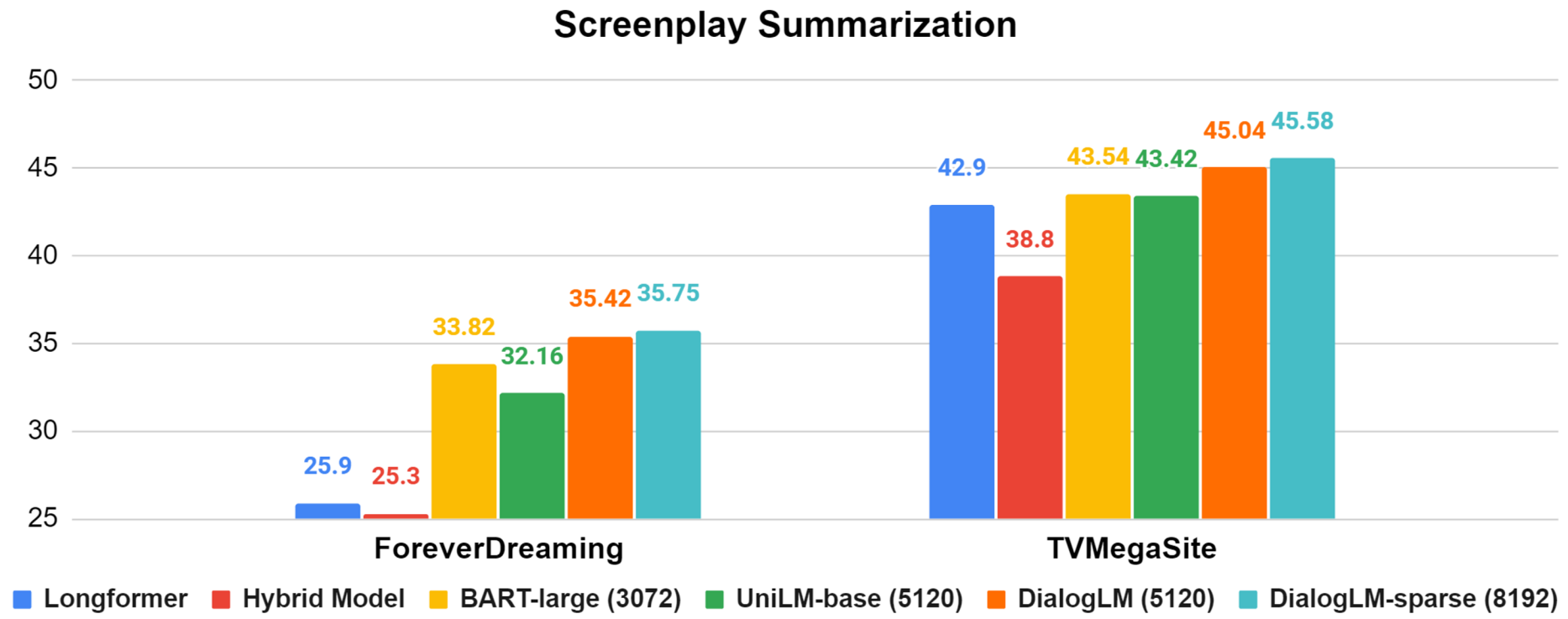
Pretraining: DialogLM

- Meeting summarization tasks



Pretraining: DialogLM

- Screenplay summarization tasks



Pretraining: HMNet

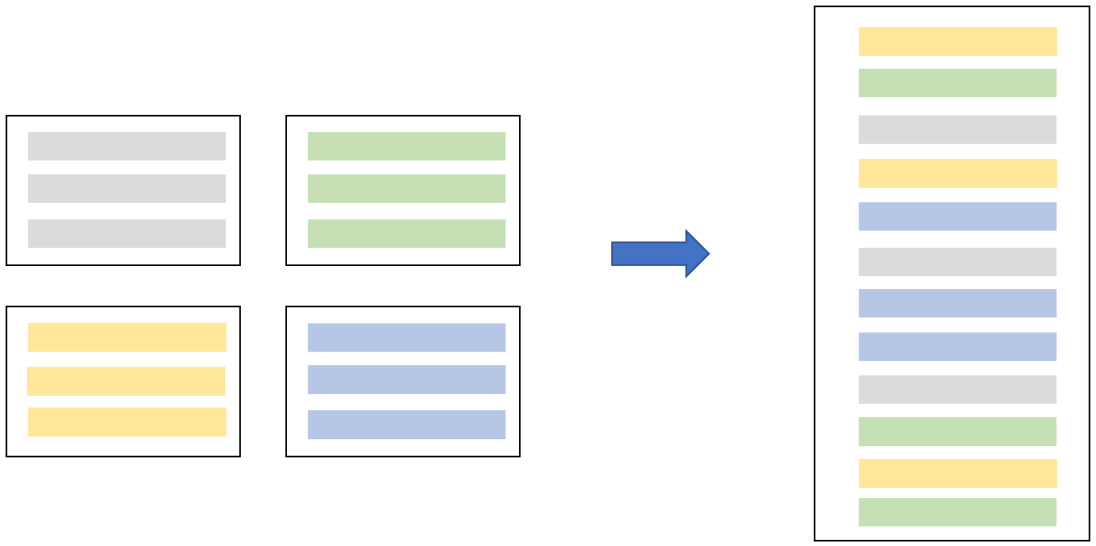
- When even pretraining data for target domain is scarce, cross-domain pretraining can help

- Given the small amount of meeting transcript data, HMNet converts news summarization data into meeting style for pre-training

[2] *A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining*
C. Zhu et al. Findings of EMNLP 2020

Pretraining: HMNet

- Cross-domain pretrain the model on news summarization datasets: CNN/DM, NYTimes and XSUM
- Given K news articles, treat each sentence from the i^{th} article as a turn from i^{th} speaker
- Randomly shuffle all turns and concatenate the turns
- The target summary is the concatenated summary for all K news articles



Model	ROUGE-1	R-2	R-SU4
	AMI		
HMNet	**53.0**	**18.6**	**24.9**
−pretrain	48.7	18.4	23.5
	ICSI		
HMNet	**46.3**	**10.6**	**19.1**
−pretrain	42.3	**10.6**	17.8

Summary of Pretraining

- Particularly useful when there's not enough in-domain large-scale labeled dialogue summarization data

Pretraining Model	When to use	Method
DialogLM	Large-scale unlabeled dialogue data	Window-based denoising
HMNet	Large-scale labeled cross-domain summarization data	Convert doc-summary data into dialog summary data

Modeling

- Each existing dialogue summarization model tackles one or more challenges specific to dialogue

- [Challenge] Long dialogue
- [Solution] Hierarchical modeling (HMNet[2]), retrieve-then-summarize (QMSum[3]), Sliding window[4]

- [Challenge] Multiple participants
- [Solution] Speaker-aware Supervised Contrastive Learning[5], Coreference-aware Summarization[6]

- [Challenge] Related Knowledge
- [Solution] Topic words and utterance structure (TGDGA[7]), Domain knowledge[8], Medical ontology (Dr. Summarize[9])

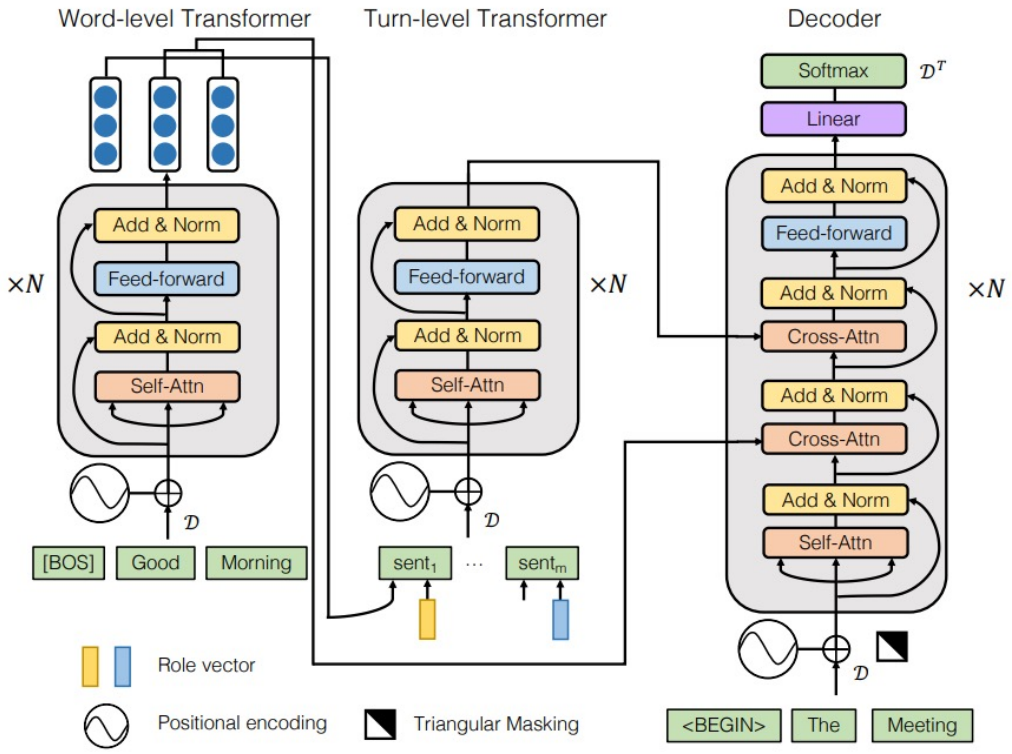
Long Dialogue

- Dialogues with transcript consisting of 1K-10K tokens are typical
- Long input is harder to fit into existing deep neural networks and also harder to summarize

- Solution
  - Use neural structure adapted to long input like hierarchical network, LongFormer[10]
  - Query-based summarization with retrieve-then-summarize

Long Dialogue: HMNet

- Encoder processes each turn at word-level and then the whole transcript at turn-level
  - The embedding of [BOS] token for each turn is used as the turn embedding for turn-level
- Decoder conducts cross attention to both word-level and turn-level embeddings
- Each speaker is assigned a personalized role vector



[2] *A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining*
C. Zhu et al. Findings of EMNLP 2020

Experiment: HMNet

Automatic Evaluation

	AMI			ICSI		
Model	ROUGE-1	R-2	R-SU4	ROUGE-1	R-2	R-SU4
Random	35.13	6.26	13.17	29.28	3.78	10.29
Template	31.50	6.80	11.40	/	/	/
TextRank	35.25	6.9	13.62	29.7	4.09	10.64
ClusterRank	35.14	6.46	13.35	27.64	3.68	9.77
UNS	37.86	7.84	14.71	31.60	4.83	11.35
Extractive Oracle	39.49	9.65	13.20	34.66	8.00	10.49
PGNet	40.77	14.87	18.68	32.00	7.70	12.46
Copy from Train	43.24	12.15	14.01	34.65	5.55	10.65
MM (TopicSeg+VFOA)*	**53.29**	13.51	/	/	/	/
MM (TopicSeg)*	51.53	12.23	/	/	/	/
HMNet	53.02	**18.57****	**24.85****	**46.28****	**10.60****	**19.12****

- Hierarchical design is particularly helpful
- Role vector depicts differences between participants

Human Evaluation

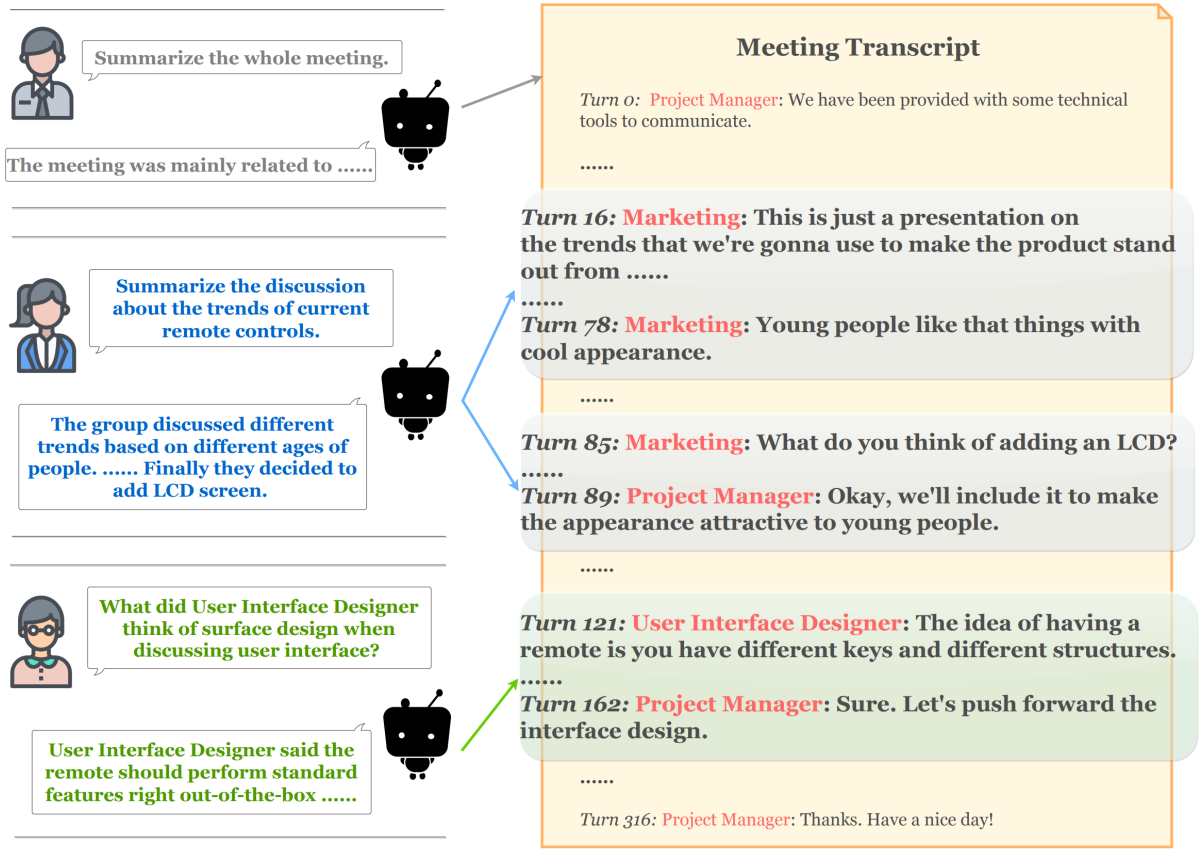
Dataset	AMI	
Source	HMNet	UNS
Readability	**4.17** (.38)	2.19 (.57)
Relevance	**4.08** (.45)	2.47 (.67)
Dataset	ICSI	
Source	HMNet	UNS
Readability	**4.24** (.20)	2.08 (.20)
Relevance	**4.02** (.55)	1.75 (.61)

Ablation Study

Model	ROUGE-1	R-2	R-SU4
	AMI		
HMNet	**53.0**	**18.6**	**24.9**
−pretrain	48.7	18.4	23.5
−role vector	47.8	17.2	21.7
−hierarchy	45.1	15.9	20.5
	ICSI		
HMNet	**46.3**	**10.6**	**19.1**
−pretrain	42.3	**10.6**	17.8
−role vector	44.0	9.6	18.2
−hierarchy	41.0	9.3	16.8

Long Dialogue: QMSum

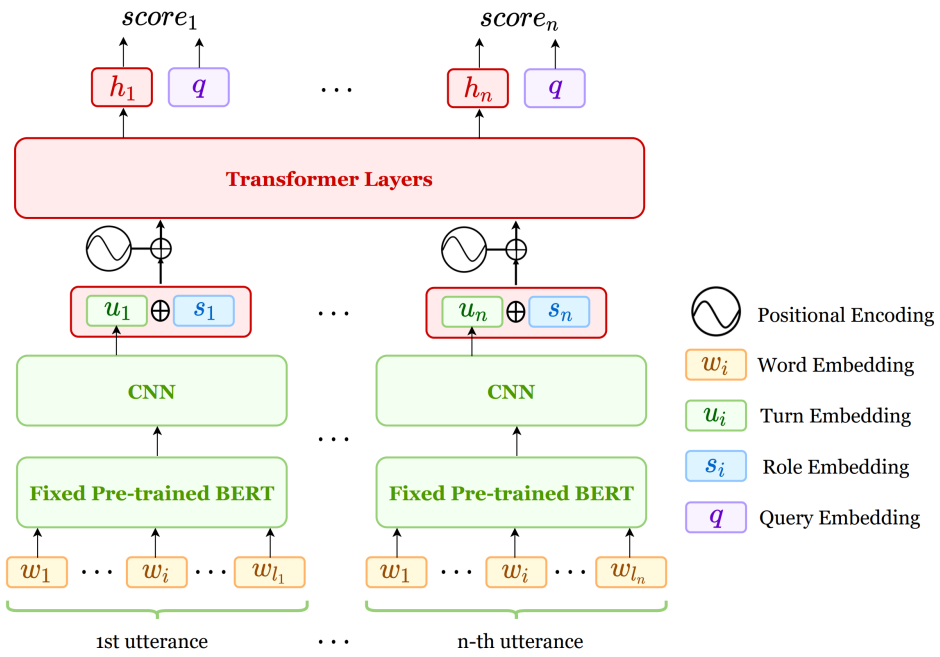
- There are multiple aspects of a meeting
- Challenging to cover all content in one single summary
- Solution: Query-based summarization
- Given the transcript and a query, focus on parts of the meeting related to the query and generate the summary answering that question



[3] *QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization.*
M. Zhong. NAACL 2021

Long Dialogue: QMSum

- Strategy: Retrieve-then-summarize
- Stage 1: Locator
- Task: locate text spans in the meeting relevant with the query
- Model 1: Pointer Network can produce multiple <start, end> pairs for text spans
- Model 2: Hierarchical ranking-based model
  - Use frozen BERT + CNN to get turn embeddings
  - Feed Turn + speaker embedding into Transformer
  - Obtain score for each turn based on its contextual embedding + query embedding
  - Train with binary cross entropy loss



Long Dialogue: QMSum

- Stage 2: Summarizer
- Task: summarize the selected text spans given the query
- Input: *<s> Query </s> Relevant Text Spans </s>”*
- Models:
  - Pointer-Generator Network
  - BART
  - HMNet

Experiments: QMSum

Locator Recall (ROUGE-L)

Models	Extracted Length			
	1/6	1/5	1/4	1/3
Random	58.86	63.20	67.56	73.81
Similarity	55.97	59.24	63.45	70.12
Pointer	61.27	65.84	70.13	75.96
Our Locator	**72.51**	**75.23**	**79.08**	**84.04**

Hierarchical ranking-based locator is the best

*: use retrieved text spans by locator

†: use gold text spans

Summarizer

Models	R-1	R-2	R-L
Random	12.03	1.32	11.76
Ext. Oracle	42.84	16.86	39.20
TextRank	16.27	2.69	15.41
PGNet	28.74	5.98	25.13
BART	29.20	6.37	25.49
PGNet*	31.37	8.47	27.08
BART*	31.74	8.53	**28.21**
HMNet*	**32.29**	**8.67**	28.17
PGNet†	31.52	8.69	27.63
BART†	32.18	8.48	28.56
HMNet†	**36.06**	**11.36**	**31.27**

Multiple Participants

- A dialogue consists of utterances from multiple participants
- The difference and interaction between speakers are vital to the understanding and summarization of a dialogue

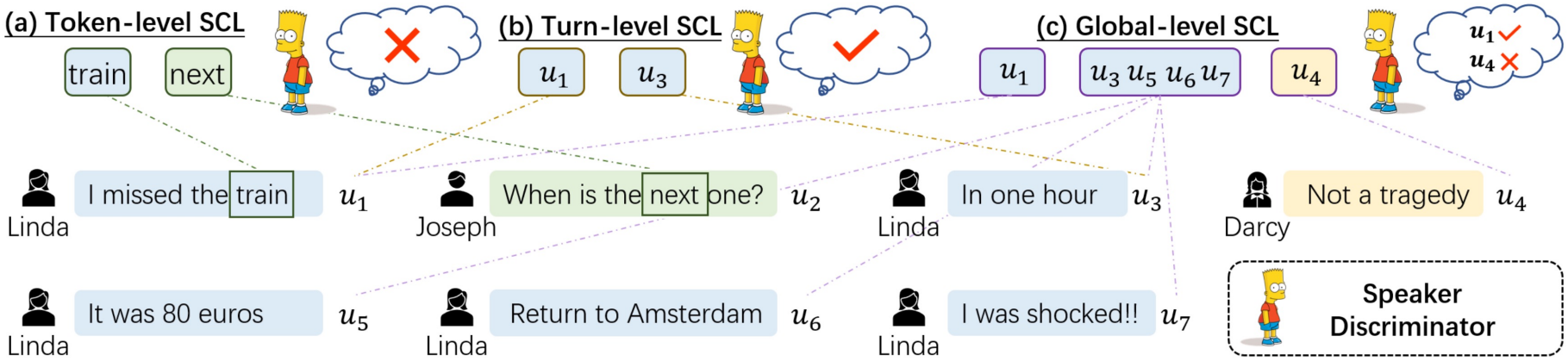
- To model the interaction between participants
  - Contrastive learning on utterances pairs from same/different speakers
  - Injecting coreference resolution information
  - Graph neural network

Multiple Participants: SCL

- Experiments show that existing dialogue models can hardly differentiate between different speakers
  - Aggregated BART embedding of K sampled tokens from utterance A and B
  - MLP classifier to judge whether A and B are from the same speaker
  - Random guess accuracy is **50%**
  - Classifier accuracy for vanilla BART: **58.1%**
  - Classifier accuracy for BART fine-tuned on SAMSUM: **60.2%**
- Teach model about the difference between different speakers' utterances
  - Supervised Contrastive Learning (SCL)
  - $(\boldsymbol{o_i}, s_i)$ denotes a sampled token / utterance's embedding and the associated speaker
  - $\mathcal{L}^+ = \sum_{\{i,j\}}^{\{s_i=s_j\}} -\log(\sigma(\boldsymbol{o_i^T o_j}))$, $\mathcal{L}^- = \sum_{\{i,j\}}^{\{s_i\neq s_j\}} -\log(\sigma(\boldsymbol{o_i^T o_j}))$
  - $\mathcal{L} = \mathcal{L}_{gen} + \lambda(\mathcal{L}^+ + \mathcal{L}^-)$

[5] *Improving Abstractive Dialogue Summarization with Speaker-Aware Supervised Contrastive Learning.*
Z. Geng et al. COLING 2022

Multiple Participants: SCL



- Token-level: sample two tokens from same/different speakers
- Turn-level: sample two turns from same/different speakers
- Global-level: sample one turn from speaker A and also one from B, comparing with the rest turns of speaker A

Experiments: SCL

Automatic evaluation

Model	SAMSum R-1	SAMSum R-2	SAMSum R-L	AMI R-1	AMI R-2	AMI R-L
PGNet (See et al., 2017)	40.08	15.28	36.63	42.60	14.01	22.62
UniLM (Dong et al., 2019; Zhu et al., 2021)	50.00	26.03	42.34	50.61	**19.33**	25.06
Multi-view BART (Chen and Yang, 2020)	53.42	27.98	49.97	-	-	-
BART+DialoGPT (Feng et al., 2021b)	53.70	28.79	50.81	-	-	-
PGN+DialoGPT (Feng et al., 2021b)	-	-	-	50.91	17.75	24.59
BART	53.01	28.05	49.89	50.67	17.18	24.96
BART + Token-level SCL task	53.85	29.21	50.94	51.03	17.23	25.21
BART + Turn-level SCL task	54.12	29.53	51.10	51.15	17.85	**25.45**
BART + Global-level SCL task	**54.22**	**29.87**	**51.35**	**51.40**	17.81	25.30

- Modeling difference between speakers can improve summary quality
- Also reduce errors such as missing a speaker in the reference summary, confusing speaker and semantic errors

Human evaluation

Model	BART	BART + Global SCL
Speaker Confusion Rate	0.100	0.067
Speaker Missing Rate	0.267	0.167
Semantic Errors Rate	0.283	0.242

Multiple Participants: Coref

- During a dialogue, participants often refer to themselves, others, concepts, objects, etc.
- Correctly understanding coreference is critical to high-quality summarization, especially to reducing factual errors

Example dialogue with annotated coreference

Max: Know any good sites to buy clothes from?
Payton: Sure :) <file_other> <file_other> <file_other>
Max: That's a lot of them!
Payton: Yeah, but they have different things so I usually buy things from 2 or 3 of them.
Max: I'll check them out. Thanks.
... ...
Max: Do u like shopping?
Payton: Yes and no.
Max: How come?
Payton: I like browsing, trying on, looking in the mirror and seeing how I look, but not always buying.
... ...
Max: So what do u usually buy?
Payton: Well, I have 2 things I must struggle to resist!
Max: Which are?
Payton: Clothes, ofc ;)
Max: Right. And the second one?
Payton: Books. I absolutely love reading!
... ...

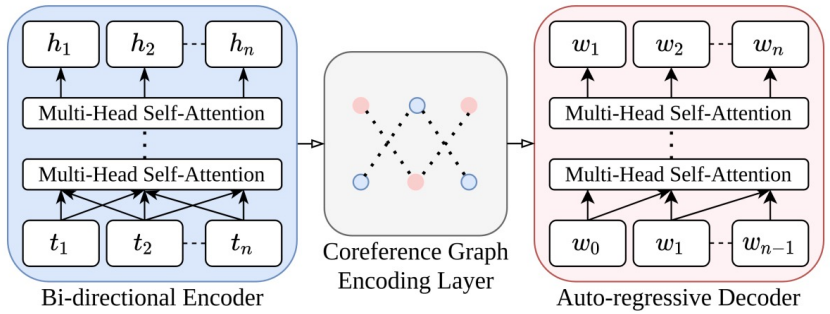
[6] *Coreference-Aware Dialogue Summarization*. Z. Liu et al. SIGDIAL 2021

Multiple Participants: Coref

- Step 1: Dialogue coreference resolution
- Apply a document coreference resolution model *Coref-SpanBERT*, obtaining coreference clusters
- Postprocessing
  - Apply model ensembling to improve accuracy
  - Assign labels to speaker role words that were not included in any cluster
  - Compare the clusters and merge those representing the same coreference chain
- Human evaluation shows that postprocessing reduces incorrect coreference assignments by ~19%

Multiple Participants: Coref

- Step 2: Use coreference information in summarization model
- Method 1: GNN
- Each entity in a cluster is a node
- Connect each entity node to its predecessor
- Each entity's initial embedding is the encoder's output H
- Multi-layer message passing to get H^G
- Linearly combine H and H^G to send to decoder

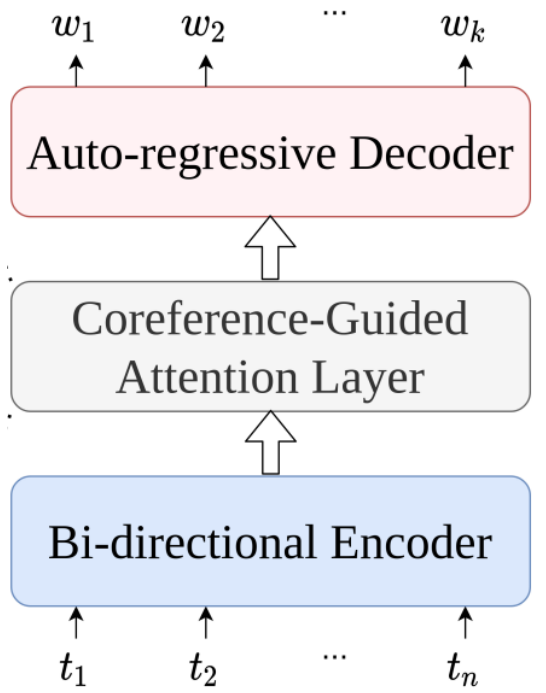


Multiple Participants: Coref

- Method 2: Coreference-Guided Attention
- Inject coreference information between encoder and decoder
- Share part of the contextual embeddings among words in the same cluster

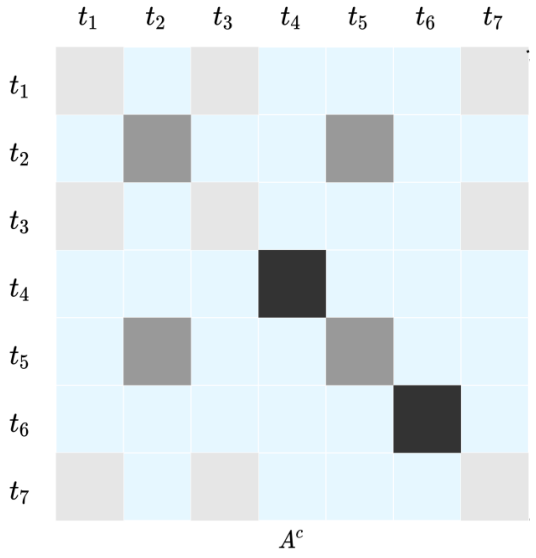
$$a_i = \sum_{j \in C^*} \frac{1}{|C^*|} h_j \text{ for } i \in C^*$$

$$h_i^A = \lambda h_i + (1 - \lambda) a_i$$



Multiple Participants: Coref

- Method 3: Coreference-Informed Transformer
- Compute attention weights A^c where each word only attends to other words in the same cluster
- In encoder's attention, select two heads whose attention weights are closest to A^c, measured by cosine similarity
- Replace the attention matrix in these two heads by A^c



Experiments: Coref

Automatic Evaluation

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
*Pointer-Generator**	40.1	-	-	15.3	-	-	36.6	-	-
*Fast-Abs-RL-Enhanced**	42.0	-	-	18.1	-	-	39.2	-	-
*DynamicConv-News**	45.4	-	-	20.6	-	-	41.5	-	-
*BART-Large**	48.2	49.3	51.7	24.5	25.1	26.4	46.6	47.5	49.5
*Multi-View BART-Large**	49.3	51.1	52.2	**25.6**	26.5	**27.4**	**47.7**	49.3	**49.9**
BART-Base	48.7	50.8	51.5	23.9	25.8	24.9	45.3	48.4	47.3
Coref-GNN	50.3	**56.1**	50.3	24.5	27.3	24.6	46.0	**50.9**	46.8
Coref-Attention	**50.9**	54.6	**52.8**	25.5	27.4	26.8	46.6	50.0	48.4
Coref-Transformer	50.3	55.5	50.9	25.1	**27.7**	25.6	46.2	**50.9**	46.9

Human Evaluation

Model	Average Scores
BART-Base	0.60
Coref-GNN	0.84
Coref-Attention	**1.16**
Coref-Transformer	0.96

Knowledge Integration

- To better summarize a dialogue, additional knowledge is needed, such as domain knowledge, utterance relationship, key entities, etc.

- Integrate these types of knowledge into summarization process
  - Graph attention
  - Attention weight computation
  - Additional term in loss

Knowledge Integration: TGDGA

- Traditional seq2seq methods struggle to handle long-distance cross-sentence dependencies in transcript, as well as maintaining relevance in generated summary

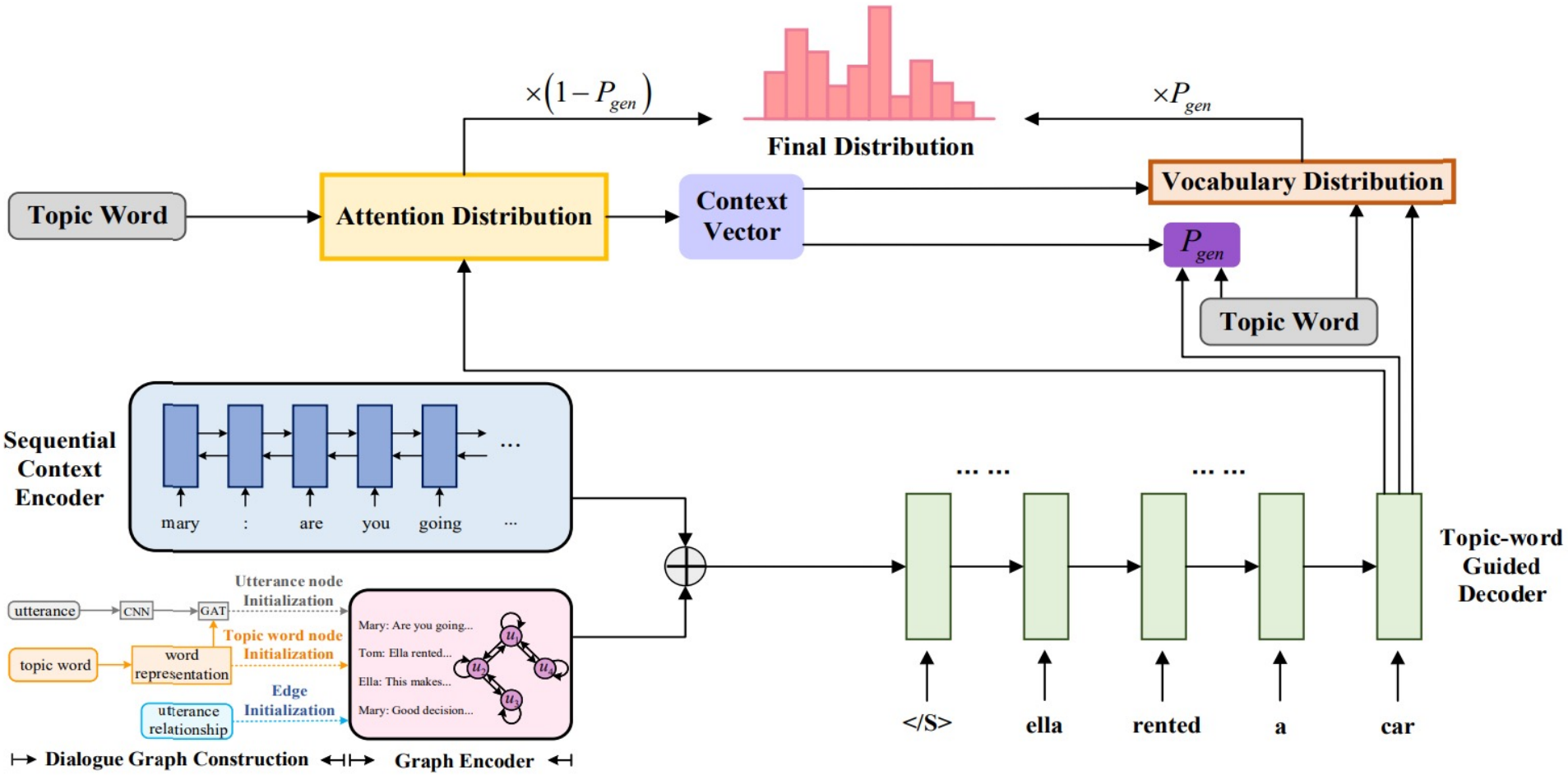
- Solution:
  - Construct a topic-word interactive graph for the dialogue
  - Utilize Graph Neural Networks (GNN) to embed graph information
  - Utilize GNN embeddings in decoder
  - Topic-word Guided Dialogue Graph Attention (**TGDGA**) network

[7] *Improving Abstractive Dialogue Summarization with Graph Structures and Topic Words*.
L. Zhao et al. COLING 2020

Knowledge Integration: TGDGA

- Step 1: Construct Graph
  - Each **topic word** is a node and each **utterance** is a node. Connect a topic word node to utterance node if the topic word is in that utterance. Connect two utterance nodes if they share at least one topic word.
- Step 2: Graph Neural Network
  - Use Masked Graph Self-Attention Layer
- Step 3: Use GNN embeddings in decoder
  - Embeddings from GNN and LSTM encoder are concatenated and fed to decoder
  - Topic word embeddings are used in coverage and pointer mechanism

Knowledge Integration: TGDGA



Experiments: TGDGA

Automatic Evaluation

Model	SAMSum Corpus			Automobile Master Corpus		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Longest-3	32.46	10.27	29.92	30.72	9.07	28.14
Seq2Seq	21.51	10.83	20.38	25.84	13.82	25.46
Seq2Seq + Attention	29.35	15.90	28.16	30.18	16.52	29.37
Transformer	36.62	11.18	33.06	36.21	11.13	34.08
Transformer + Separator	37.27	10.76	32.73	37.43	11.87	34.97
LightConv	33.19	11.14	30.34	34.68	12.41	31.62
DynamicConv	33.79	11.19	30.41	34.72	12.45	31.86
DynamicConv + Separator	33.69	10.88	30.93	34.41	12.38	31.22
Pointer Generator	38.55	14.14	34.85	39.17	15.39	34.76
Pointer Generator + Separator	40.88	15.28	36.63	39.23	15.42	34.53
Fast Abs RL	40.96	17.18	39.05	39.82	15.86	36.03
Fast Abs RL Enhanced	41.95	18.06	39.23	40.13	16.17	36.42
TGDGA (ours)	43.11	19.15	40.49	42.98	17.58	38.11

Human Evaluation

Dataset	Model	Relevance	Readability
SAMSum	Pointer Generator + Separator	2.36	4.25
	Fast Abs RL Enhanced	2.67	4.73
	TGDGA (ours)	2.91	4.86
Automobile Master	Pointer Generator + Separator	2.41	4.18
	Fast Abs RL Enhanced	2.59	4.35
	TGDGA (ours)	2.88	4.62

Knowledge Integration: Dr. Summarize

- Medical summarization has unique challenges
  - Transcripts contain many domain-specific terms
  - Factuality of summary is particularly important
- Solution
  - Utilize information from **medical ontologies**
  - Handle **negation**
  - **Encourage copying** and penalize generation from dictionary
- Base model: seq2seq pointer-generator network with coverage loss

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t \qquad c^t = \sum_{t'=0}^{t-1} a^{t'} \qquad \text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t)$$

[9] *Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures.*

A. Joshi et al. Findings of EMNLP 2020

Knowledge Integration: Dr. Summarize

Medical knowledge

- Use medical knowledge from Unified Medical Language Systems (UMLS)
- The one-hot vector m^t encodes presence of UMLS medical concepts in both source transcript and target summary
- Apply m^t in attention and loss only during training

$$e_i^t = v^t \tanh(W_h h_i + W_s s_t + w_c c_i^t + w_m m_i^t + w_n n_i^t + b_{attn})$$

Medical knowledge Negation

Negation

- The one-hot vector n_i^t encodes whether the t-th source word is a negative word, e.g., no, not, doesn't. Use n_i^t in attention.

Factuality

- Encourage copy and penalize generation from vocabulary by adding δp_{gen} into loss function

Experiments: Dr. Summarize

Models	Metrics			Doctor Evaluation			
	Negation F1	Concept F1	ROUGE-L F1	Model	Baseline	Both	None
2M-BASE	70.1±0.8	69.1±1.3	52.6±0.9	-	-	-	-
2M-PGEN	67.3±3.3	72.8±0.8	55.4±0.9	37.1%	18.5 %	38.9 %	5.3%
2M-PGEN-NEG	72.2±3.6	70.9±2.2	53.5±0.7	37.7%	22.7%	34.1%	5.4%
3M-PGEN-NEG-CONCEPT	78.0±4.2	70.6±1.4	55.2±1.2	26.9%	25.7%	42.5%	4.2%

3M means 2M + predicting special token [NO]

- Both automatic and doctor evaluations show gains after applying generation penalty (PGEN), negation (NEG) and medical term coverage (CONCEPT)

Summary of Modeling

- Mostly based on seq2seq models
- Address problems specific to dialogue summarization

Problem Addressed	Work	Method
Long Dialogue	HMNet[2]	Hierarchical network
	QMSum[3]	Retrieve-then-summarize, Hierarchical network
Multiple Participants	SCL[5]	Constrastive loss on utterance speaker
	Coref[6]	Injecting coreference information via GNN and attention
Knowledge Integration	TGDGA[7]	Apply GNN to entity-utterance graph
	Dr. Summarize[9]	Leverage medical ontologies, handle negation, encourage copying

More information about dialogue summarization: [11] *A Survey on Dialogue Summarization.* X. Feng et al. IJCAI 2022

Reference

[1] DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. M. Zhong et al. AAAI 2022

[2] A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. C. Zhu et al. Findings of EMNLP 2020

[3] QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. M. Zhong et al. NAACL 2021

[4] A Sliding-Window Approach to Automatic Creation of Meeting Minutes. J. J. Koay et al. NAACL: Student Research Workshop, 2021

[5] Improving Abstractive Dialogue Summarization with Speaker-Aware Supervised Contrastive Learning. Z. Geng et al. COLING 2022

[6] Coreference-Aware Dialogue Summarization. Z. Liu et al. SIGDIAL 2021

[7] Improving Abstractive Dialogue Summarization with Graph Structures and Topic Words. L. Zhao et al. COLING 2020

[8] How Domain Terminology Affects Meeting Summarization Performance. J. J. Koay et al. COLING 2020

[9] Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures. A. Joshi et al. Findings of EMNLP 2020

[10] Longformer: The Long-Document Transformer. I. Beltagy et al. arXiv, 2020.

[11] A Survey on Dialogue Summarization. X. Feng et al. IJCAI 2022