Tutorial: Summarization of Dialogues and Conversations At Scale

Diyi Yang[1], Chenguang Zhu[2]
[1]Stanford University [2]Microsoft Cognitive Services Research

What is This Tutorial About

- How to summarize dialogues
- Application, Datasets, Models, Challenges

Who We Are

Diyi Yang

Assistant professor in the Computer Science Department at Stanford University

Research: dialogue summarization, learning with limited text data, and computational social science.

Chenguang Zhu

Principal Research Manager in Microsoft Cognitive Services Research Group

Research: dialogue summarization, knowledge graph, prompt learning and multimodal learning.

Schedule

Local time (CEST)	Content	Presenter
14:15-14:45	Introduction to Conversation Summarization	Diyi Yang
14:45-15:15	Conversation Structures and Evaluation	Diyi Yang
15:15-15:30	Break	
15:30-16:50	Pretraining and Models	Chenguang Zhu
16:50-17:00	Conclusion and Challenges	Chenguang Zhu

Where can I find more information?

Talk slides, reading materials, schedule:
https://github.com/zcgzcgzcg1/EACL2023_Tutorial_Dialogue_Summarization

Tutorial Proposal
https://aclanthology.org/2023.eacl-tutorials.3.pdf

Disclaimer: This tutorial is presenters' own opinions

- To access mentioned models + datasets, please refer to corresponding licensing information
- We're not promoting the use of any particular model and/or datasets
- There are slides / figures borrowed from respective papers
- This tutorial is by no means exhaustive: we've tried our best to include relevant materials

Outline

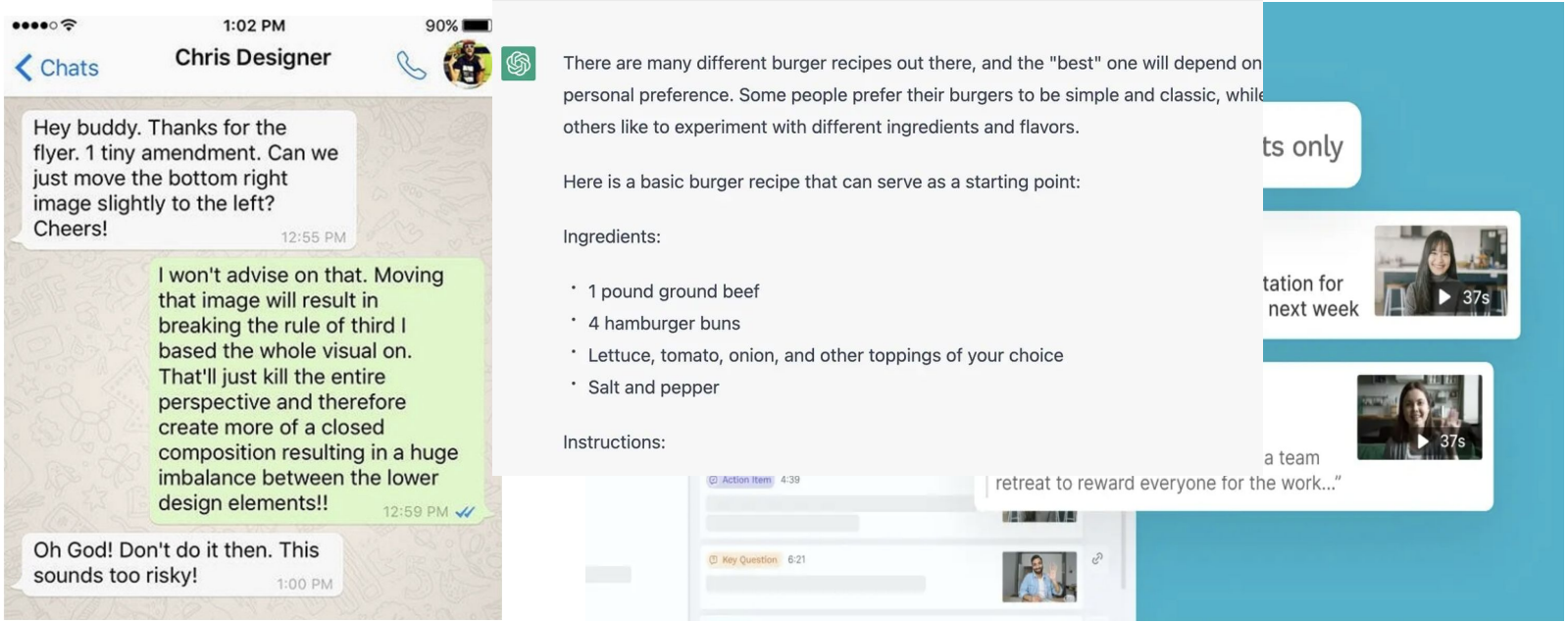
1. Introduction to Conversation Summarization
2. Conversation Structures and Evaluation
3. Pre-training and Models
4. Conclusion and Future Directions

Outline

1. Introduction to Conversation Summarization
2. Conversation Structures and Evaluation
3. Pre-training and Models
4. Conclusion and Future Directions

Communications via **textual dialogues** are increasing.

Communications via **textual dialogues** are increasing.

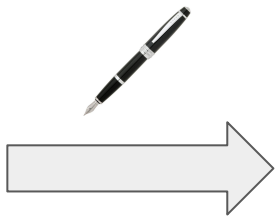


https://mobiletrans.wondershare.com/whatsapp-tips/export-whatsapp-chat-iphone.html
https://techcrunch.com/2023/03/31/reads-ai-powered-summary-feature-squeezes-a-meeting-into-a-two-minute-clip/

How to **process** and **organize** interactions into **concise and structured** summaries becomes technically and socially important!

Abstractive Dialogue Summarization

Hannah: Hey, do you have Betty's number?
Amanda: Lemme check
Amanda: Sorry, can't find it.
Amanda: Ask Larry
Amanda: He called her last time we were at the park together
Hannah: I don't know him well
Amanda: Don't be shy, he's very nice
Hannah: If you say so..
Hannah: I'd rather you texted him
Amanda: Just text him ?
Hannah: Urgh.. Alright
Hannah: Bye
Amanda: Bye bye



Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.

Dialog Types in Linguistics

Mainly around conversation analysis and pragmatics.

A large body of work investigates speech acts (Searle, 1969; Grice, 1975), i.e. dialogues are decomposed into individual turns and their communicative intents

Some around typology of dialogues (Walton and Krabbe (1995)

Are We Summarizing the Right Way?

The Walton & Krabbe Model

Six basic dialogue types:

Persuasion, Negotiation, Inquiry, Deliberation, Information-seeking, and Eristics.

Three additional mixed types:

Debate (Persuasion and Eristics), Committee meeting (mainly Deliberation), and Socratic Dialogue

	Persuasion	Negotiation	Inquiry	Deliberation	Information-seeking	Eristics
Initial situation	Conflicting points of view (POVs)	Conflict of interests & need for cooperation	General ignorance	Need for action	Personal Ignorance	Conflict & antagonism
Main goal	Resolution of such conflicts by verbal means	Making a deal	Growth of knowledge & agreement	Reach a decision	Spreading knowledge and revealing positions	Reaching a (provisional) accommodation in a relationship
Participants' aim	Persuade the other(s)	Get the best out of it for oneself	Find a "proof" or destroy one	Influence out-come	Gain, pass on, show, or hide personal knowledge	Strike the other party & win in the eyes of onlookers
Side benefits	Develop and reveal positions, Build up confidence, Influence onlookers, Add to prestige	Agreement, Build up confidence, Reveal position Influence onlookers, Add to prestige	Add to prestige, Gain experience, Raise funds	Agreement, Develop & reveal positions, Add to prestige, Vent emotions	Agreement, Develop & reveal positions, Add to prestige, Vent emotions	Agreement, Develop & reveal positions, Gain experience, Amusement, Add to prestige, Vent emotions

	Persuasion	Negotiation	Inquiry	Deliberation	Information-seeking	Eristics
Initial situation	Conflicting points of view (POVs)	Conflict of interests & need for cooperation	General ignorance	Need for action	Personal Ignorance	Conflict & antagonism
Main goal	Resolution of such conflicts by verbal means	Making a deal	Growth of knowledge & agreement	Reach a decision	Spreading knowledge and revealing positions	Reaching a (provisional) accommodation in a relationship
Participants' aim	Persuade the other(s)	Get the best out of it for oneself	Find a "proof" or destroy one	Influence out-come	Gain, pass on, show, or hide personal knowledge	Strike the other party & win in the eyes of onlookers
Side benefits	Develop and reveal positions, Build up confidence, Influence onlookers, Add to prestige	Agreement, Build up confidence, Reveal position Influence onlookers, Add to prestige	Add to prestige, Gain experience, Raise funds	Agreement, Develop & reveal positions, Add to prestige, Vent emotions	Agreement, Develop & reveal positions, Add to prestige, Vent emotions	Agreement, Develop & reveal positions, Gain experience, Amusement, Add to prestige, Vent emotions
Summary items	POVs, Resolutions, Disagreements, Positions, Arguments, Winners/Losers, Controversies	Final deal, Initial interests, Winners/Losers, Evolution of deal, Arguments	Initial inquiry, Gained/new knowledge, Reached agreement, (Line of) arguments, Mentioned facts	Decision, Initial need for action, Positions of speakers, Evolution of decision, Winners/Losers, Emotions	Initial problem, Solution, Positions, Emotions	Initial conflict, Resolution/agreement, Winners/Losers, Arguments, Emotions
NLP targets	Topics, Stances, Decisions, Arguments, Emotions, Sentiment	Decisions, Stances, Topic tracking, arguments	Topics, Knowledge, Decisions, Arguments, Keyfacts	Decisions, Topics, Stances, Arguments, Topic tracking, Emotions	Topics, Action items, Decisions, Stances, Emotions	Topics, Action items, Decisions, Arguments, Emotions

Applications and Datasets

Applications & Datasets

- **Meeting Summarization**
  - ICSI (Janin et al., 2003)

https://www1.icsi.berkeley.edu/Speech/mr/mtgrcdr.html

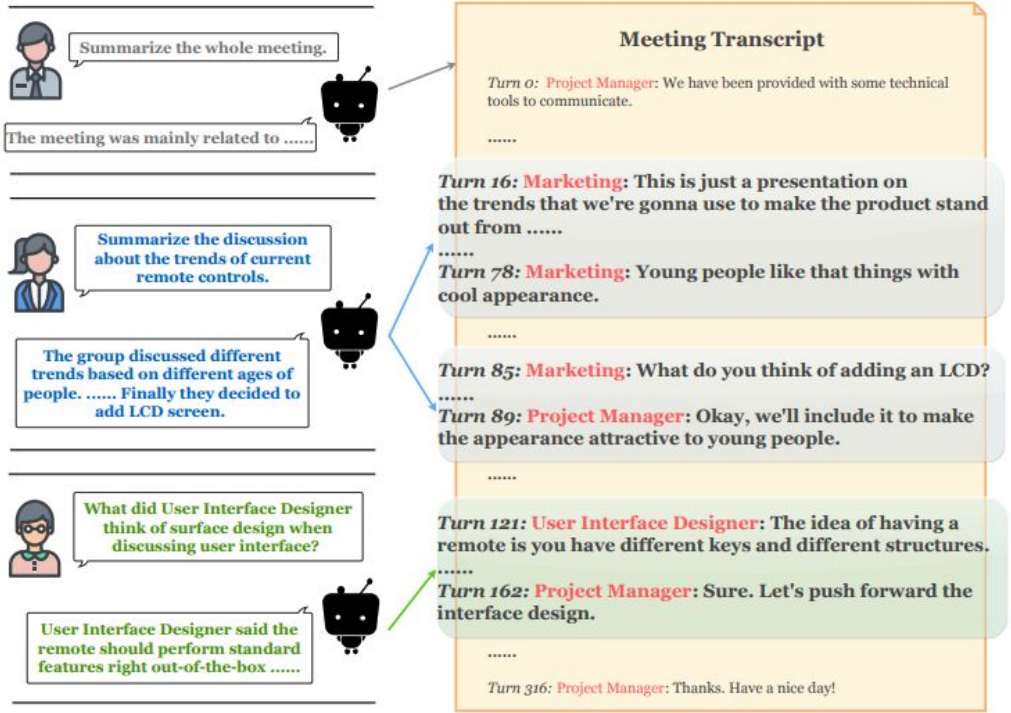
Applications & Datasets

- Meeting Summarization
  - ICSI (Janin et al., 2003)
  - AMI (Carletta et al., 2005)
    - a multi-modal data set consisting of 100 hours of meeting recordings

Annotation: orthographic transcription, annotations for many different phenomena (dialog acts, head movement etc.).

Applications & Datasets

- Meeting Summarization
  - ICSI (Janin et al., 2003)
  - AMI (Carletta et al., 2005)
  - QMSum (Zhong et al., 2021)



Examples of query-based meeting summarization task. Users are interested in different facets of the meeting. In this task, a model is required to summarize the contents that users are interested in and query

Applications & Datasets

- Meeting Summarization
- Chat Summarization
  - SAMSum (Gliwa et al., 2019)
  - GupShup (Mehnaz et al., 2021)
  - DialogSum (Chen et al., 2021)

Dialogue

Blair: Remember we are seeing the wedding planner after work
Chuck: Sure, where are we meeting her?
Blair: At Nonna Rita's
Chuck: Can I order their seafood tagliatelle or are we just having coffee with her? I've been dreaming about it since we went there

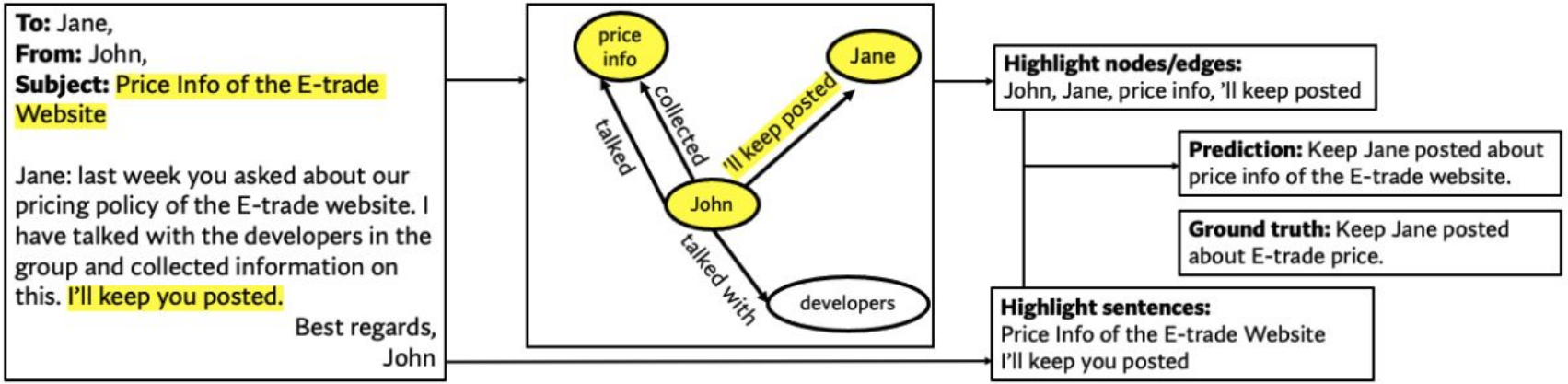
Leon: kya tujeh abhi tak naukari nahi mili?
Arthur: nahi bro, abhi bhi unemployed :D
Leon: hahaha, LIVING LIFE
Arthur: mujeh yeh bahot acha lagta hai, dopahar ko jagata hoon, sports dekhta hoon - ek aadmi ko aur kya chahiye?
Leon: a paycheck? ;)
Arthur: mean mat bano ...
Leon: but seriously, mere dosth ke company mein ek junior project manager offer hai, tujeh interest hai?
Arthur: sure thing, tere pass details hai?
Leon: <file_photo>

English Summary: *Arthur is still unemployed. Leon sends him a job offer for junior project manager position. Arthur is interested.*

Datasets	Lan. style	Domain	Scenario	Dialogs	Data size	#Tokens/dial.	#Tokens/turn
AMI	spoken	single	meeting	137	100hrs (video)	4,757	16.5
SAMSum	written	multiple	online	16,369	1.5M (token)	94	8.4
DIALOGSUM	spoken	multiple	daily life	13,460	1.8M (token)	131	13.8

Applications & Datasets

- Meeting Summarization
- Chat Summarization
- Email Thread Summarization
  - EMAILSUM (Zhang et al., 2021a)



Applications & Datasets

- Meeting Summarization
- Chat Summarization
- Email Thread Summarization
- Customer Service Summarization
  - CSDS (Lin et al., 2021)
  - TODSum (Zhao et al., 2021)

Applications & Datasets

- Meeting Summarization
- Chat Summarization
- Email Thread Summarization
- Customer Service Summarizat
- Medical Dialogue Sum
  - CMC (Song et al., 2020)

	Role	Utterance	Translation	Tag
Conv.	P	胆碱能性荨麻疹怎么治疗	How to treat cholinergic urticaria and measles	PD
	D	这种情况多长时间了？用什么治疗过？	How long has this condition last? What treatment have you used?	OT
	P	好长时间了，之前治疗过，中西药都吃过就没治好	It has been a long time. I have taken both Chinese and Western medicine, but it is not working.	OT
	D	主要是避免诱因。胆碱能性荨麻疹要保持身体凉爽、避免出汗、避免精神紧张、进食热饮或酒精饮料等。	You need to avoid triggers of cholinergic urticaria. Keep your body cool and avoid sweating, mental stress, hot drink, alcoholic beverages, etc.	DT
	P	那怎么样能根治呢	How can it be cured?	OT
	D	目前医疗上，没有明确的根治方法。	At present, there is no clear cure for this disease.	OT
	D	内服药物之外还可以中药外洗.这个方法也有一定的效果。蚕砂、苦参、芒硝、白矾、荆芥准备二十克，把这些药一起煎了进行外洗，一天二次。	In addition to taking medicines, you can also wash the skin with Chinese medicine, which has some effect. Use the decoction of Silkworm litter, Sophora flavescens, Glauber's salt, alum, and Nepeta, 20 grams for each, to wash your skin twice a day.	DT
SUM1		胆碱能性荨麻疹怎么治疗	How to treat cholinergic urticaria and measles	
SUM2	A	胆碱能性荨麻疹要保持身体凉爽，避免出汗，避免精神紧张、进食热饮或酒精饮料等。内服药物之外还可以通过中药进行外洗。这个方法也有一定的效果。蚕砂、苦参、芒硝、白矾、荆芥准备二十克,把这些药一起煎了进行外洗，一天二次。	You need to avoid triggers of cholinergic urticaria. keep your body cool and avoid sweating, mental stress, hot drink, alcoholic beverages, etc. In addition to taking medicines, you can also wash the skin with Chinese medicine, which has some effect. Use the decoction of Silkworm litter, Sophora flavescens, Glauber's salt, alum, and Nepeta, 20 grams for each, to wash your skin twice a day.	
	B	口服脱敏药物，同时避免诱因。胆碱能性荨麻疹要保持身体凉爽，避免出汗，避免精神紧张、进食热饮或酒精饮料等。	You need to take desensitization drugs and avoid triggers. You should keep the body cool, avoid sweating, mental stress, hot drink, alcoholic beverages, etc.	

Applications & Datasets

- Meeting Summarization
- Chat Summarization
- Email Thread Summarization
- Customer Service Summarization
- Medical Dialogue Summarization
- Media Summarization
  - MEDIASUM (Zhu et al., 2021)
  - SumTitles (Malykh et al., 2020)

Interview transcripts from NPR and CNN and employ the overview and topic descriptions as summaries

movie scripts as the main source of data for corpus construction

Major datasets for dialogue summarization

Name	Domain	Language
ICSI [Janin *et al.*, 2003]	Meeting	English
AMI [Carletta *et al.*, 2005]		English
QMSum [Zhong *et al.*, 2021]		English
SAMSum [Gliwa *et al.*, 2019]	Chat	English
GupShup [Mehnaz *et al.*, 2021]		Code-Mix
CSDS [Lin *et al.*, 2021]	Customer Service	Chinese
TODSum [Zhao *et al.*, 2021]		English
TWEETSUMM [Feigenblat *et al.*, 2021]		English
CRD3 [Rameshkumar and Bailey, 2020]	TV Show	English
[Song *et al.*, 2020]	Medical	Chinese
SumTitles [Malykh *et al.*, 2020]	Movie	English
MEDIASUM [Zhu *et al.*, 2021]	Interview	English
DIALOGSUM [Chen *et al.*, 2021]	Spoken	English
EMAILSUM [Zhang *et al.*, 2021a]	Email	English
ForumSum [Khalman *et al.*, 2021]	Forum	English
ConvoSumm [Fabbri *et al.*, 2021]	Mix	English

Feng, Xiachong, Xiaocheng Feng, and Bing Qin. "A survey on dialogue summarization: Recent advances and new frontiers." *arXiv preprint arXiv:2107.03175* (2021).

Outline

1. Introduction to Conversation Summarization
2. Conversation Structures and Evaluation
3. Pre-training and Models
4. Conclusion and Future Directions

Unique Aspects about Conversations

- Informal
- Verbose and repetitive
- Back channeling, reconfirmations
- Hesitations
- Speaker interruptions
- Multiple Speakers
- ...

Utilizing Structures for Conversation Summarization

- Transfering document summarization model (Gliwa et al., 2019)
- Adopting hierarchical models (Zhao et al., 2019; Zhu et al. 2020)
- Incorporating conversation structures (Liu et al., 2019; Li et al., 2019; Chen and Yang, 2020, Chen and Yang 2021; Liu et al., 2019b; Feng et al., 2020c)
- Learning from Human Feedback (Chen et al. 2022)

Transfering Document Summarization Model

Model	Train data	Separator	R-1	R-2	R-L
LONGEST-3 baseline	n/a	n/a	32.46	10.27	29.92
Pointer Generator	dialogues	no	38.55	14.14	34.85
Pointer Generator	dialogues	yes	40.08	15.28	36.63
Fast Abs RL	dialogues	no	40.96	17.18	39.05
Fast Abs RL Enhanced	dialogues	no	41.95	18.06	39.23
Transformer	dialogues	no	36.62	11.18	33.06
Transformer	dialogues	yes	37.27	10.76	32.73
LightConv	dialogues	no	33.19	11.14	30.34
DynamicConv	dialogues	no	33.79	11.19	30.41
DynamicConv	dialogues	yes	33.69	10.88	30.93
LightConv + GPT-2 emb.	dialogues	no	41.81	16.34	37.63
DynamicConv + GPT-2 emb.	dialogues	no	41.79	16.44	37.54
DynamicConv + GPT-2 emb.	dialogues	yes	41.54	16.29	37.07

ROUGE scores on SamSUM using document summarization models.

Gliwa, Bogdan, et al. "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization."

Incorporating Conversation Structures

- There are different types of structures in conversations.
- Utilizing these diverse structures help understanding and summarizing conversations.

Topic Segmentation

- One single conversation may cover multiple topics during interactions
- *“greetings → invitation → party details → rejection”*

Xing, Linzi, and Giuseppe Carenini. "Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring”
Chen, Jiaao, and Diyi Yang. "Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization."

Topic Segmentation

Turns	Dialogue Text
Turn-1: A: For how long should the liability insurance coverage remain in effect? **Turn-2**: B: As long as the registration of your vehicle remains valid. **Turn-3**: A: Does this apply for motorcycles too? **Turn-4**: B: There are some exceptions for motorcycles.	
Turn-5: A: Regarding the name on my vehicle registration application and the one on the Insurance Identification Card, do they need to be the same? **Turn-6**: B: yes, the names must match in both documents.	
Turn-7: A: Can I submit copies or faxes of my Insurance identification card to the DMV? **Turn-8**: B: yes, you can. But take into consideration that the card will be rejected if the DMV barcode reader can not scan the barcode.	

Xing, Linzi, and Giuseppe Carenini. "Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring"
Chen, Jiaao, and Diyi Yang. "Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization."

Conversation Stages

- Conversations develop following certain patterns.

Conversation	Topic View	Stage View
James: Hey! I have been thinking about you :)	*Greetings*	*Openings*
Hannah: Oh, that's nice ;)		
James: What are you up to?	*Today's plan*	*Intentions*
Hannah: I'm about to sleep		
James: I miss u. I was hoping to see you		
Hannah: Have to get up early for work tomorrow	*Plan for tomorrow*	*Discussion*
James: What about tomorrow?		
Hannah: To be honest I have plans for tomorrow evening		
James: Oh ok. What about Sat then?	*Plan for Saturday*	
Hannah: Yeah. Sure I am available on Sat		
James: I'll pick you up at 8?	*Pick up time*	
Hannah: Sounds good. See you then		*Conclusion*

Althoff, Tim, Kevin Clark, Jure Leskovec. "Large-scale analysis of counseling conversations: An application of natural language processing to mental health."
Chen, Jiaao, Diyi Yang. "Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization."

Dialog Acts

- Dialogue acts reflect the the effect of an utterance on the context.

Multi-Party Dialogue	Dialogue Act
A: mm-hmm .	Backchannel
B: mm-hmm .	Backchannel
C: then , these are some of the remotes which are different in shape and colour , but they have many buttons .	Inform
C: so uh sometimes the user finds it very difficult to recognise which button is for what function and all that .	Inform
D: so you can design an interface which is very simple , and which is user-friendly .	Inform
D: even a kid can use that .	Inform
A: so can you got on t t uh to the next slide .	Suggest
Summary: alternative interface options	

Goo, Chih-Wen, and Yun-Nung Chen. "Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts."

Coreferences

- People usually refer to each other, mention others' names or use coreference in their messages.

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!
... ...

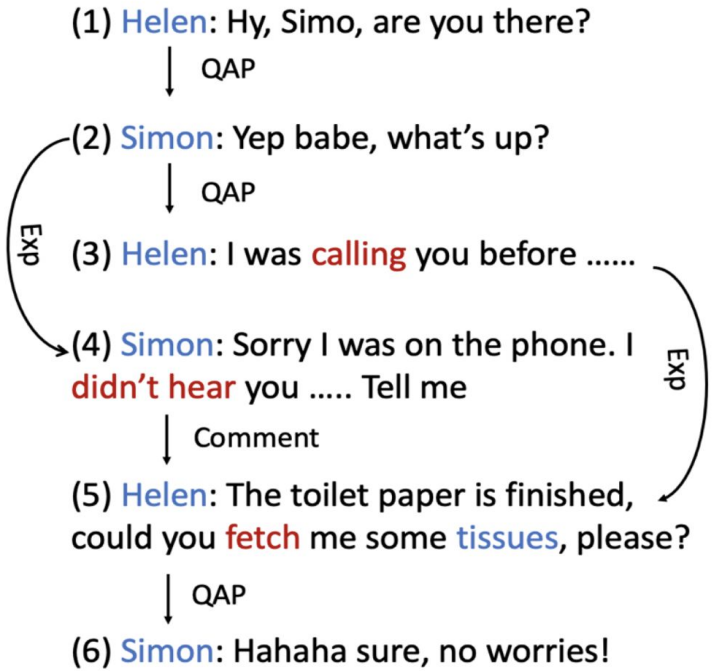
Base Model: *Payton* is looking for good places to buy clothes. He usually buys things from 2 or 3 of them. He likes browsing and trying on clothes. *Max* likes reading books.

Coreference-Aware Model: Max will check out some good places to buy clothes. Payton likes browsing, trying on, looking in the mirror and seeing how she looks. Payton loves reading.

Liu, Zhengyuan, Ke Shi, and Nancy F. Chen. "Coreference-aware dialogue summarization."

Discourse Relations

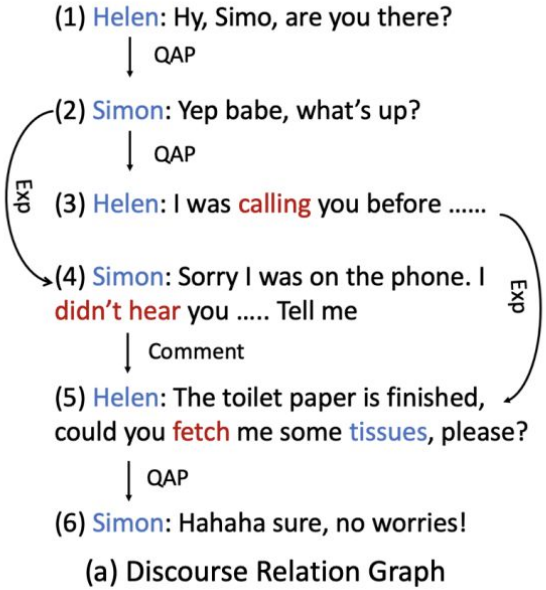
- Utterances in conversations are related with each others within the context of discourse.



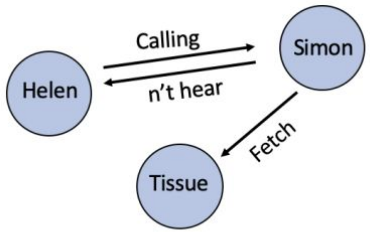
Chen, Jiaao, and Diyi Yang. "Structure-aware abstractive conversation summarization via discourse and action graphs."

Action Mentions

- "*WHO-DOING-WHAT*" represents the action information in conversations.



(a) Discourse Relation Graph

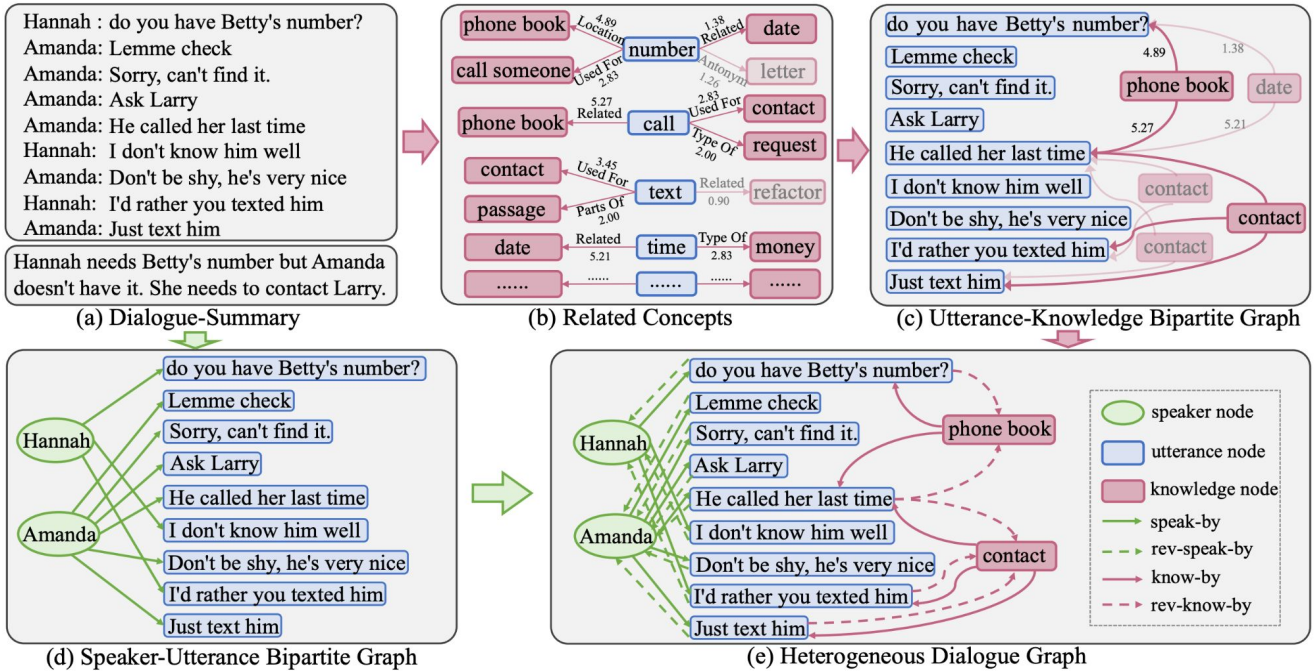


(b) Action Graph

Chen, Jiaao, and Diyi Yang. "Structure-aware abstractive conversation summarization via discourse and action graphs."

Commonsense

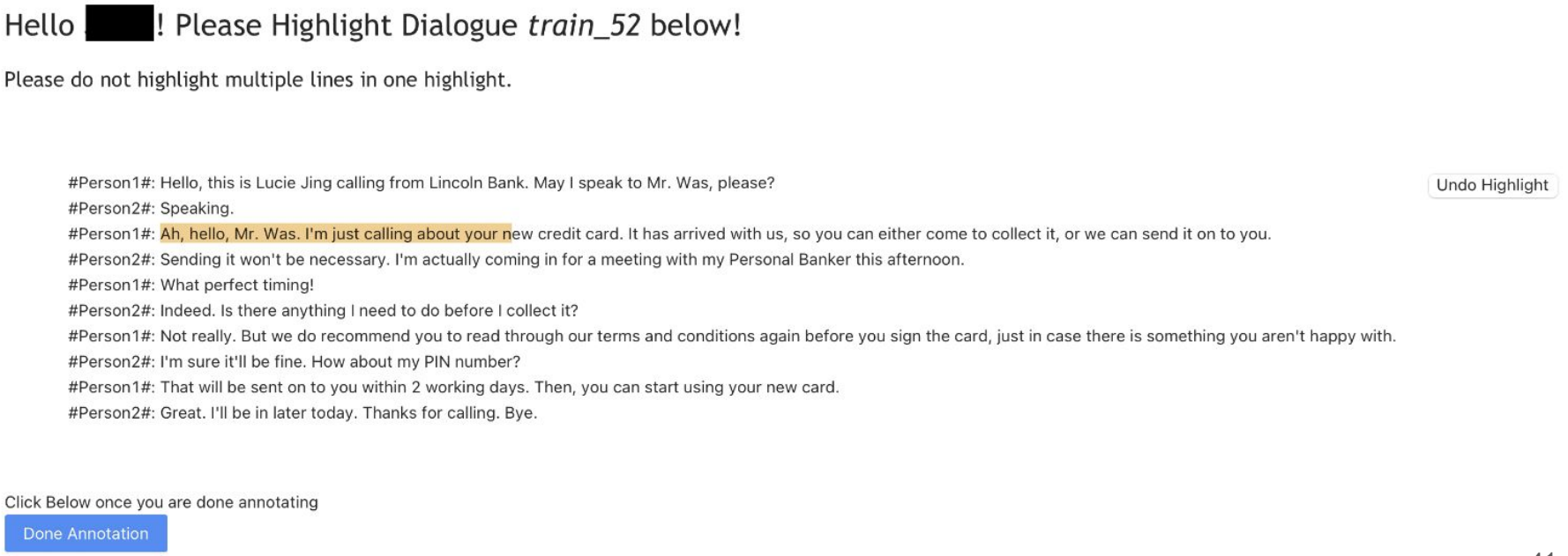
- Commonsense to better capture conversation nuances



Feng, Xiachong, Xiaocheng Feng, Bing Qin. "Incorporating commonsense knowledge into abstractive dialogue summarization via graph networks."

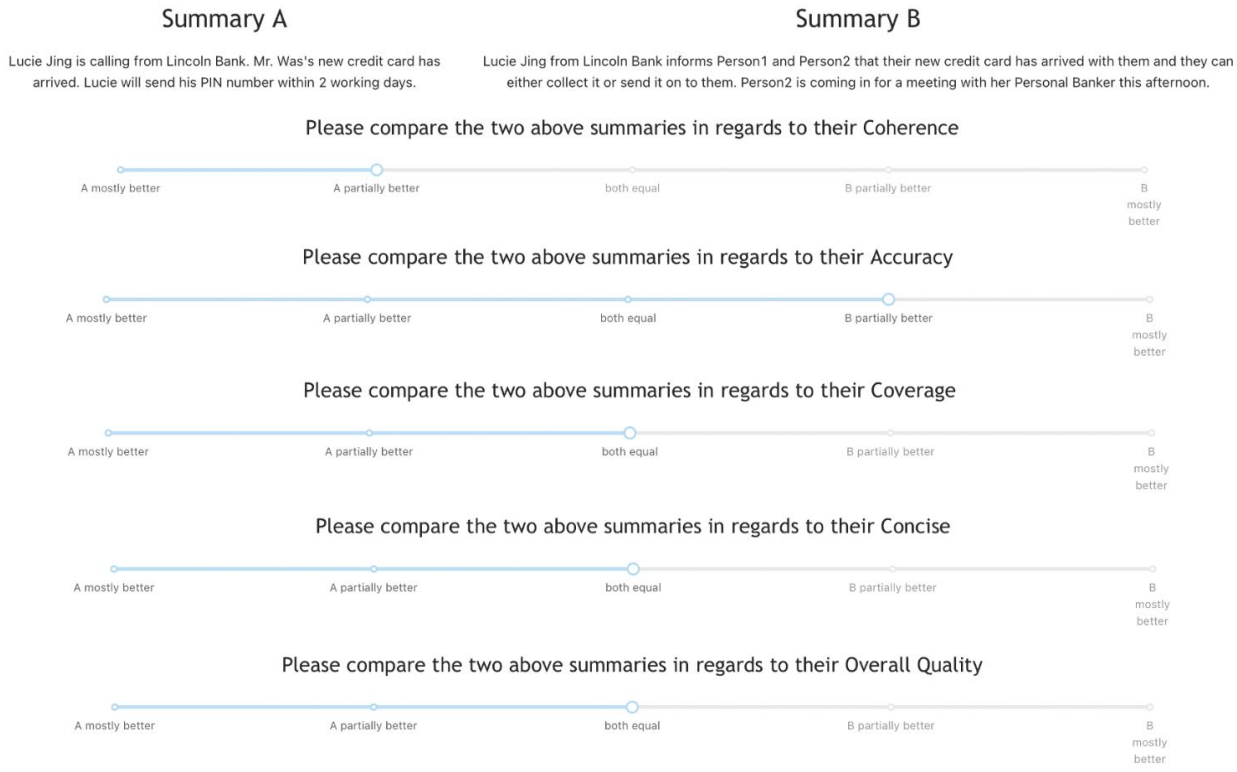
Human Feedback towards Summarizations

- Local Feedback

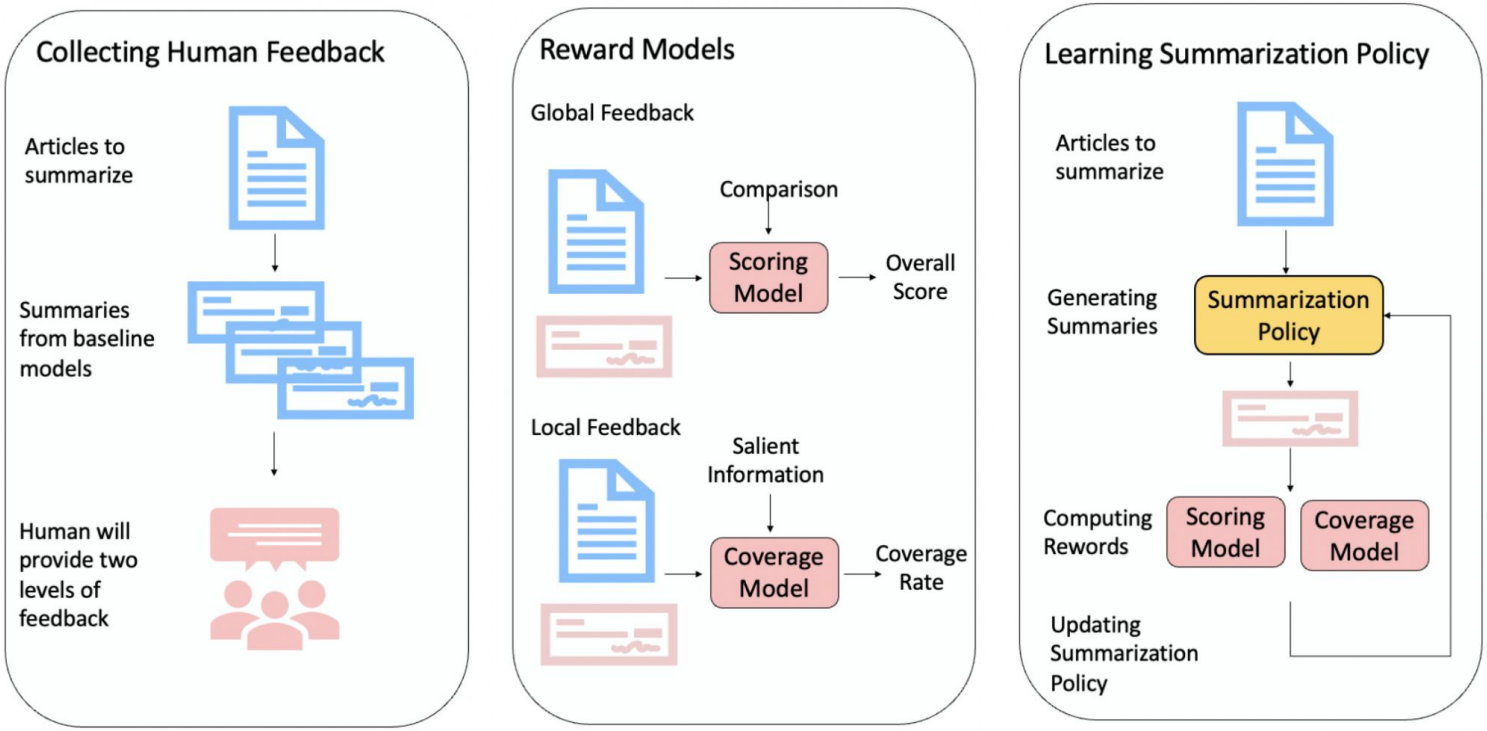


Human Feedback towards Summarizations

- Global Feedback



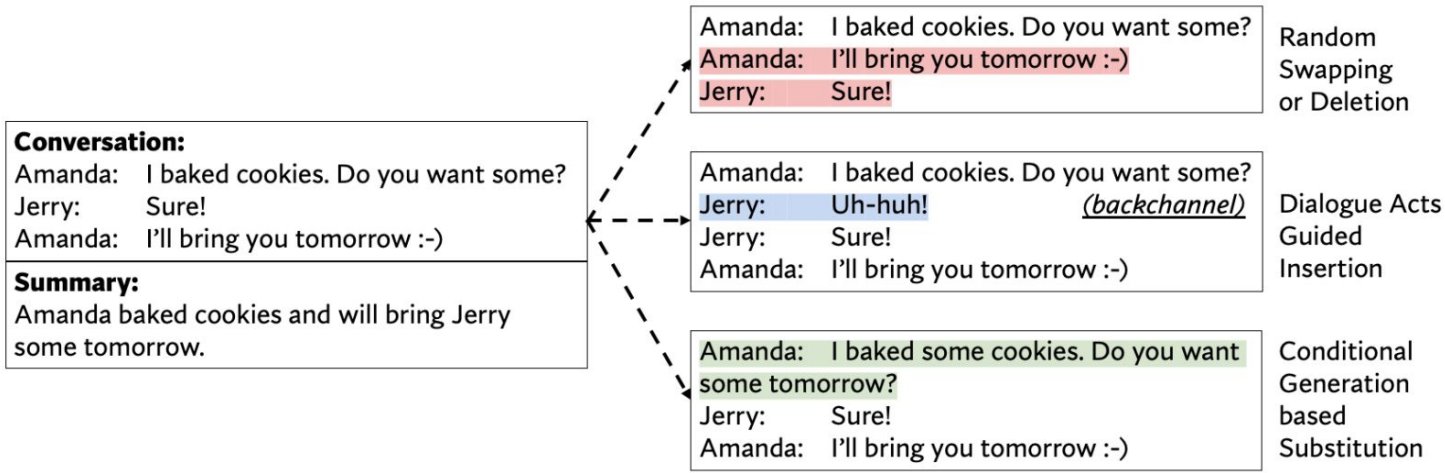
Learning from Human Feedback



Chen, Jiaao, Mohan Dodda, and Diyi Yang. "Human-in-the-loop Abstractive Dialogue Summarization."

What if limited labeled conversations?

- Data Augmentation
  - Generating novel conversation-summary pairs (Chen and Yang, 2021)



Chen, Jiaao, and Diyi Yang. "Simple conversational data augmentation for semi-supervised abstractive dialogue summarization."

What if limited labeled conversations?

- Semi-supervised Learning
  - Incorporating unlabeled conversations (Chen and Yang, 2021; Park and Lee, 2022; Ahmad et al, 2023)
    - Consistency Regularization
    - Self-training

Before model, let’s talk about Evaluation first

Automatic Evaluation

- ROUGE scores (Lin, 2004)
  - Word overlaps
- PRISM (Thompson and Post, 2020)
  - Making use of models pre-trained on paraphrase classification
- MoverScore/BERTScore/BARTScore (Zhao et al, 2019;Zhang et al, 2020;Yuan, 2021)
  - Utilizing contextualized representations

Automatic Evaluation

- ChatGPT

Score the following news summarization given the corresponding news with respect to fluency with one to five stars, where one star means "disfluency" and five stars means "perfect fluency". Note that fluency measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences.

News: [a news article]
Summary: [one generated summary]
Stars:

Wang, Jiaan, et al. "Is chatgpt a good nlg evaluator? a preliminary study."

ChatGPT based Evaluation

Metrics	Coherence			Relevance			Consistency			Fluency		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	0.167	0.160	0.126	0.326	0.359	0.252	0.160	0.224	0.130	0.115	0.158	0.094
ROUGE-2	0.184	0.174	0.139	0.290	0.327	0.219	0.187	0.246	0.155	0.159	0.185	0.128
ROUGE-L	0.128	0.102	0.099	0.311	0.342	0.237	0.115	0.189	0.092	0.105	0.141	0.084
BERTScore	0.283	0.310	0.211	0.311	0.346	0.243	0.110	0.152	0.090	0.192	0.209	0.158
MoverScore	0.159	0.167	0.118	0.318	0.371	0.244	0.157	0.224	0.127	0.129	0.176	0.105
PRISM	0.249	0.258	0.196	0.212	0.232	0.163	0.345	0.352	0.285	0.254	0.264	0.205
BARTScore	0.322	0.345	0.250	0.264	0.290	0.197	0.311	0.321	0.256	0.248	0.260	0.203
BARTScore+CNN	0.448	0.458	0.342	0.356	0.369	0.273	0.382	0.422	0.315	0.356	0.407	0.292
BARTScore+CNN+Para	0.424	0.442	0.325	0.313	0.364	0.241	0.401	0.487	0.332	**0.378**	**0.448**	0.311
ChatGPT	**0.470**	**0.484**	**0.403**	**0.428**	**0.454**	**0.374**	**0.419**	**0.517**	**0.389**	0.353	0.415	**0.329**

Spearman correlation (Spear.) correlation, Pearman (Pear.) correlation and Kendall's Tau (Kend.) of different aspects compared to human evaluation

Wang, Jiaan, et al. "Is chatgpt a good nlg evaluator? a preliminary study."

Fine-grained Evaluation for ConvSum?

- Broader Coverage

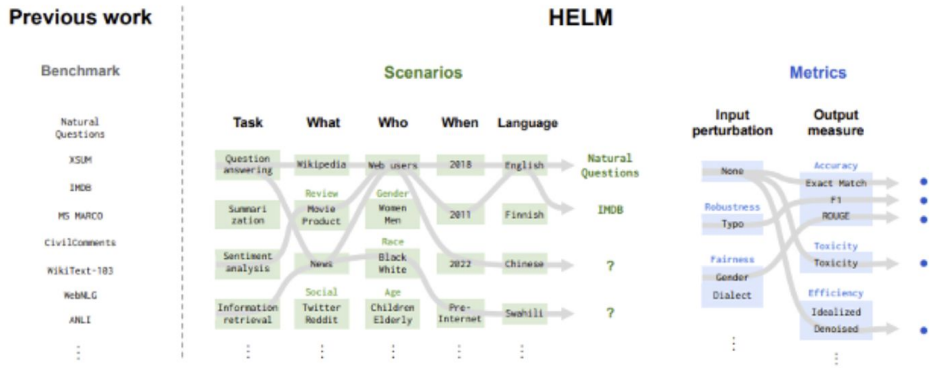


Fig. 2. **The importance of the taxonomy to HELM.** Previous language model benchmarks (e.g. SuperGLUE, EleutherAI LM Evaluation Harness, BIG-Bench) are collections of datasets, each with a standard task framing and canonical metric, usually accuracy (*left*). In comparison, in HELM we take a top-down approach of first explicitly stating what we want to evaluate (i.e. scenarios and metrics) by working through their underlying structure. Given this stated taxonomy, we make deliberate decisions on what subset we implement and evaluate, which makes explicit what we miss (e.g. coverage of languages beyond English).

Liang, Percy, et al. "Holistic evaluation of language models." *arXiv preprint arXiv:2211.09110* (2022).

Fine-grained Evaluation for ConvSum?

- Multi-metric

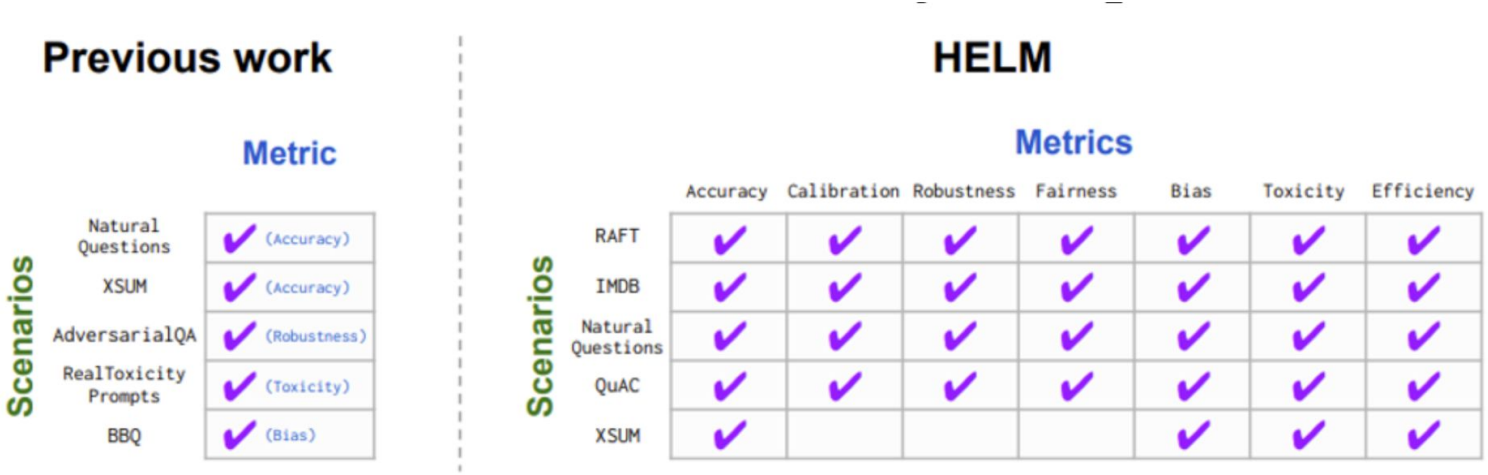
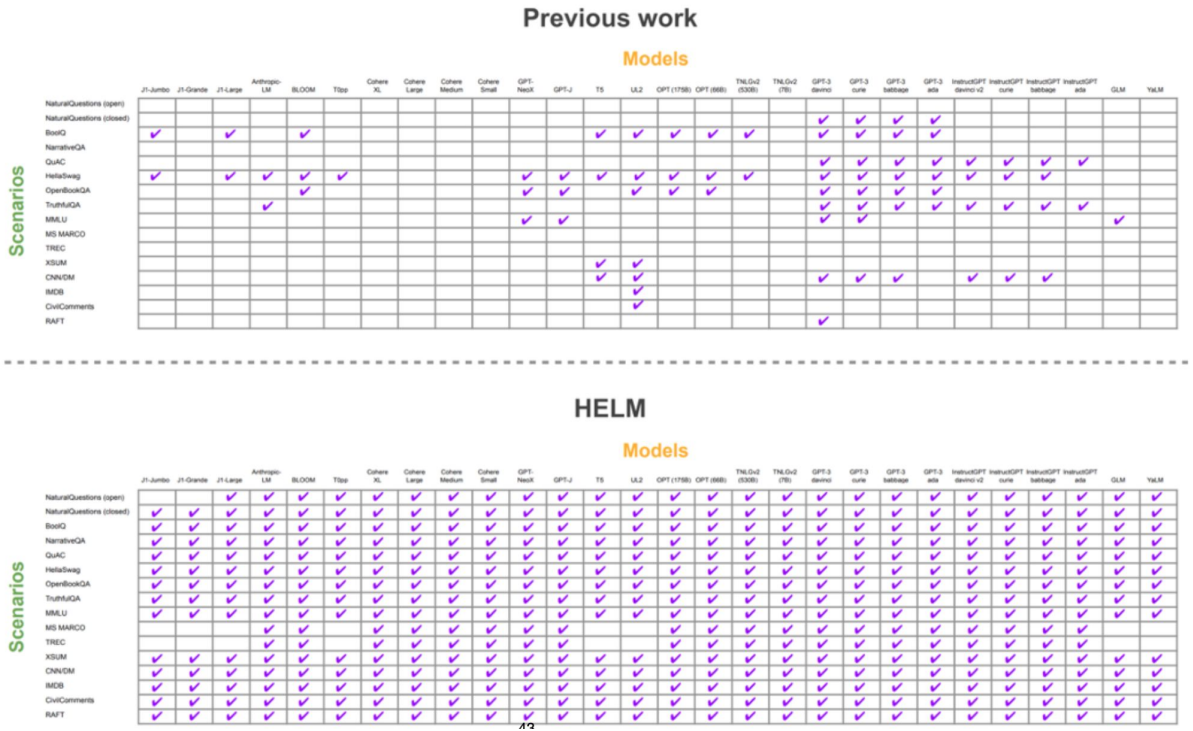


Fig. 3. **Many metrics for each use case.** In comparison to most prior benchmarks of language technologies, which primarily center accuracy and often relegate other desiderata to their own bespoke datasets (if at all), in HELM we take a multi-metric approach. This foregrounds metrics beyond accuracy and allows one to study the tradeoffs between the metrics.

Liang, Percy, et al. "Holistic evaluation of language models." *arXiv preprint arXiv:2211.09110* (2022).

Fine-grained Evaluation for ConvSum?

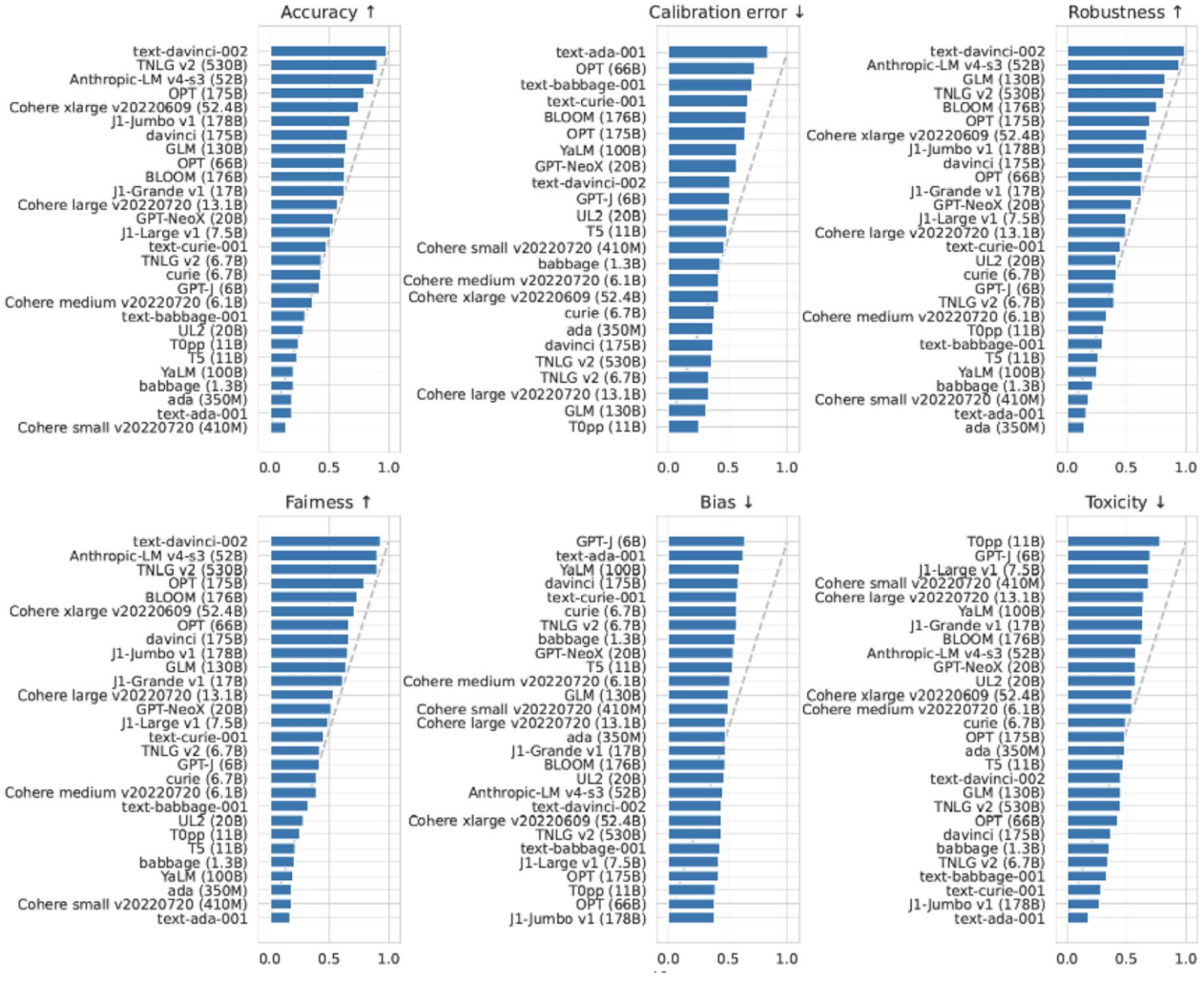
- Standardization



Desiderate/Metrics

Venue	Desiderata
ACL, EMNLP, NAACL, LREC …	accuracy, bias, environmental impact, explainability, fairness, interpretability, linguistic plausibility, robustness sample efficiency, toxicity, training efficiency
SIGIR	accuracy, bias, explainability, fairness, inference efficiency, privacy, security, user experience/interaction
NeurIPS, ICML, ICLR, …	accuracy, fairness, interpretability, privacy, robustness, sample efficiency, theoretical guarantees, training efficiency uncertainty/calibration, user experience/interaction
AAAI	accountability, accuracy, bias, causality, creativity, emotional intelligence, explainability, fairness, interpretability memory efficiency, morality, privacy, robustness, sample efficiency, security, theoretical guarantees, transparency trustworthiness, uncertainty/calibration, user experience/interaction
COLT, UAI, AISTATS	accuracy, causality, fairness, memory efficiency, privacy, sample efficiency, theoretical guarantees, training efficiency
The Web Conference (WWW), ICWSM	accessibility, accountability, accuracy, bias, credibility/provenance, fairness, inference efficiency, legality, privacy, reliability robustness, security, transparency, trustworthiness, user experience/interaction
FAccT	causality, explainability, fairness, interpretability, legality, oversight, participatory design, privacy, security transparency, user experience/interaction
WSDM	accountability, accuracy, credibility/provenance, explainability, fairness, inference efficiency, interpretability

Category	Desiderata
Requires knowledge of how model was created	causality, environmental impact, linguistic plausibility, memory efficiency, participatory design, privacy sample efficiency, training efficiency, theoretical guarantees
Requires the model have specific structure	credibility/provenance, explainability
Requires more than blackbox access	interpretability
Require knowledge about the broader system	maintainability, reliability, security, transparency
Requires knowledge about the broader social context	accessibility, accountability, creativity, emotional intelligence, legality, morality, oversight trustworthiness, user experience/interaction
Satisfies our conditions (i.e. none of the above)	accuracy, bias, fairness, inference efficiency, robustness, toxicity, uncertainty/calibration


Accuracy ↑
text-davinci-002
TNLG v2 (530B)
Anthropic-LM v4-s3 (52B)
OPT (175B)
Cohere xlarge v20220609 (52.4B)
J1-Jumbo v1 (178B)
davinci (175B)
GLM (130B)
OPT (66B)
BLOOM (176B)
J1-Grande v1 (17B)
Cohere large v20220720 (13.1B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
text-curie-001
TNLG v2 (6.7B)
curie (6.7B)
GPT-J (6B)
Cohere medium v20220720 (6.1B)
text-babbage-001
UL2 (20B)
T0pp (11B)
T5 (11B)
YaLM (100B)
babbage (1.3B)
ada (350M)
text-ada-001
Cohere small v20220720 (410M)
0.0
0.5
1.0
Calibration error ↓
text-ada-001
OPT (66B)
text-babbage-001
text-curie-001
BLOOM (176B)
OPT (175B)
YaLM (100B)
GPT-NeoX (20B)
text-davinci-002
GPT-J (6B)
UL2 (20B)
T5 (11B)
Cohere small v20220720 (410M)
babbage (1.3B)
Cohere medium v20220720 (6.1B)
Cohere xlarge v20220609 (52.4B)
curie (6.7B)
ada (350M)
davinci (175B)
TNLG v2 (530B)
TNLG v2 (6.7B)
Cohere large v20220720 (13.1B)
GLM (130B)
T0pp (11B)
0.0
0.5
1.0
Robustness ↑
text-davinci-002
Anthropic-LM v4-s3 (52B)
GLM (130B)
TNLG v2 (530B)
BLOOM (176B)
OPT (175B)
Cohere xlarge v20220609 (52.4B)
J1-Jumbo v1 (178B)
davinci (175B)
OPT (66B)
J1-Grande v1 (17B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
Cohere large v20220720 (13.1B)
text-curie-001
UL2 (20B)
curie (6.7B)
GPT-J (6B)
TNLG v2 (6.7B)
Cohere medium v20220720 (6.1B)
T0pp (11B)
text-babbage-001
T5 (11B)
YaLM (100B)
babbage (1.3B)
Cohere small v20220720 (410M)
text-ada-001
ada (350M)
0.0
0.5
1.0
Faimess ↑
text-davinci-002
Anthropic-LM v4-s3 (52B)
TNLG v2 (530B)
OPT (175B)
BLOOM (176B)
Cohere xlarge v20220609 (52.4B)
OPT (66B)
davinci (175B)
J1-Jumbo v1 (178B)
GLM (130B)
J1-Grande v1 (17B)
Cohere large v20220720 (13.1B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
text-curie-001
TNLG v2 (6.7B)
GPT-J (6B)
curie (6.7B)
Cohere medium v20220720 (6.1B)
text-babbage-001
UL2 (20B)
T0pp (11B)
T5 (11B)
babbage (1.3B)
YaLM (100B)
ada (350M)
Cohere small v20220720 (410M)
text-ada-001
0.0
0.5
1.0
Bias ↓
GPT-J (6B)
text-ada-001
YaLM (100B)
davinci (175B)
text-curie-001
curie (6.7B)
TNLG v2 (6.7B)
babbage (1.3B)
GPT-NeoX (20B)
T5 (11B)
Cohere medium v20220720 (6.1B)
GLM (130B)
Cohere small v20220720 (410M)
Cohere large v20220720 (13.1B)
ada (350M)
J1-Grande v1 (17B)
BLOOM (176B)
UL2 (20B)
Anthropic-LM v4-s3 (52B)
text-davinci-002
Cohere xlarge v20220609 (52.4B)
TNLG v2 (530B)
text-babbage-001
J1-Large v1 (7.5B)
OPT (175B)
T0pp (11B)
OPT (66B)
J1-Jumbo v1 (178B)
0.0
0.5
1.0
Toxicity ↓
T0pp (11B)
GPT-J (6B)
J1-Large v1 (7.5B)
Cohere small v20220720 (410M)
Cohere large v20220720 (13.1B)
YaLM (100B)
J1-Grande v1 (17B)
BLOOM (176B)
Anthropic-LM v4-s3 (52B)
GPT-NeoX (20B)
UL2 (20B)
Cohere xlarge v20220609 (52.4B)
Cohere medium v20220720 (6.1B)
curie (6.7B)
OPT (175B)
ada (350M)
T5 (11B)
text-davinci-002
GLM (130B)
TNLG v2 (530B)
OPT (66B)
davinci (175B)
babbage (1.3B)
TNLG v2 (6.7B)
text-babbage-001
text-curie-001
J1-Jumbo v1 (178B)
text-ada-001
0.0
0.5
1.0

Fine-grained Evaluation for ConvSum?

- Move away from using a single metric for performance evaluation.
- Evaluate social bias and efficiency.
- Consider how to aggregate multiple metrics.

Fine-grained Evaluation

- When evaluating on multiple metrics, scores are typically averaged to obtain a single score

Metric Weights	Accuracy	Throughput	Memory	Fairness	Robustness	Dynascore
DeBERTa default params (dynateam)	69.54	7.41	5.71	91.97	75.70	38.83
RoBERTa default params (dynateam)	69.07	9.23	4.82	90.94	74.82	38.61
ALBERT default params (dynateam)	67.29	9.60	2.18	89.94	74.12	37.72
T5 default params (dynateam)	67.16	7.10	10.62	91.89	73.47	37.53
BERT default params (dynateam)	64.82	9.39	4.13	92.11	66.38	36.36
Majority Baseline (dynateam)	32.41	77.33	1.15	100.00	100.00	22.78
FastText default params (dynateam)	31.29	73.94	2.20	83.23	69.14	21.13

Dynamic metric weighting in the DynaBench natural language inference task leaderboard

Challenges in Dialogue Summarization

- Informal Language Use

Bella: It's valentine's day! 😜😜😜
Aria: For sombody without *bf* today is *kinda* miserable day 😂😂😂
...

Challenges in Dialogue Summarization

- Informal Language Use
- Multiple Participants

Jeff: Do you know guys anything ...
Vladimir: the most important is ...
Tanya: and they will completely ...
Donald: yeah, mostly gas and oil.
...

Challenges in Dialogue Summarization

- Informal Language Use
- Multiple Participants
- Multiple Turns

Marty: Hiya, I have a favour to ask.
Marty: Can you pick up Marcel ...
...
(16 turns)

Challenges in Dialogue Summarization

- Informal Language Use
- Multiple Participants
- Multiple Turns
- Referral & Coreference

Phil: Good evening Deana!
...
Deana: ... belong your Cathreen!
Phil: No. *She* says *they* aren't *hers*.
...
Phil: Where did *you* find *them*?
...

Challenges in Dialogue Summarization

- Informal Language Use
- Multiple Participants
- Multiple Turns
- Referral & Coreference
- Repetition & Interruption

Greg: Well, could you pick him up?
Besty: *What if I can't?*
Greg: *Besty?*
Besty: *What if I can't?*
Greg: *Can't you, really?*
Besty: *I can't. ...*
...

Challenges in Dialogue Summarization

- Informal Language Use
- Multiple Participants
- Multiple Turns
- Referral & Coreference
- Repetition & Interruption
- Negation & Rhetorical

Sam: I *don't* think he likes me.
Cathy: Of course he likes you.
Sam: How do u know. *He's not* ...
Cathy: He's looking at u when u *don't* see
Sam: *Really? U sure?*
...

Challenges in Dialogue Summarization

- Informal Language Use
- Multiple Participants
- Multiple Turns
- Referral & Coreference
- Repetition & Interruption
- Negation & Rhetorical
- Role & Language Change

Margaret: … maybe we could meet on 17th?
Evans: … I won't also be 17th.
Margaret: OK, I get it.
Evans: But we could meet 14th, if you like?
Margaret: I'm not sure …
…

Challenges in Dialogue Summarization

- Informal Language Use
- Multiple Participants
- Multiple Turns
- Referral & Coreference
- Repetition & Interruption
- Negation & Rhetorical
- Role & Language Change
- Language Variations

Country	Total English speakers
🌍 **World** 🌏	**1,179,874,130**
🇺🇸 United States 🇺🇸	316,107,532
🇮🇳 India 🇮🇳	128,539,090
🇵🇰 Pakistan 🇵🇰	115,044,691
🇳🇬 Nigeria 🇳🇬	103,198,040

Outline

1. Introduction to Conversation Summarization
2. Conversation Structures and Evaluation
3. Pre-training and Models
4. Conclusion and Future Directions

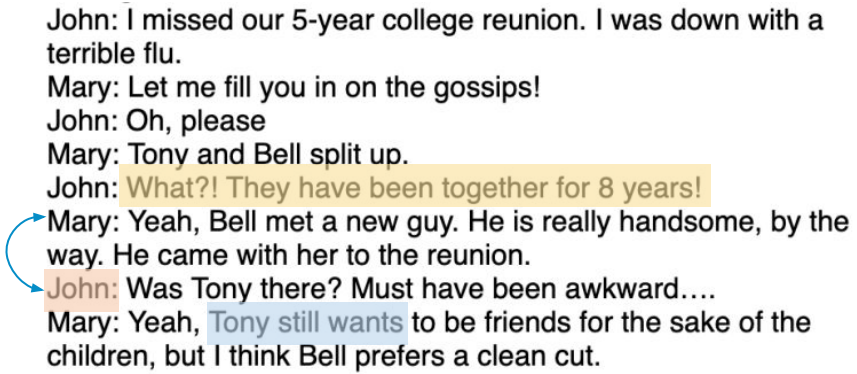
Pretraining

- 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, GPT-x etc.
- Existing pre-trained LM are primarily for documents, not proper for dialogues

Designing Pre-training Tasks for Dialogues

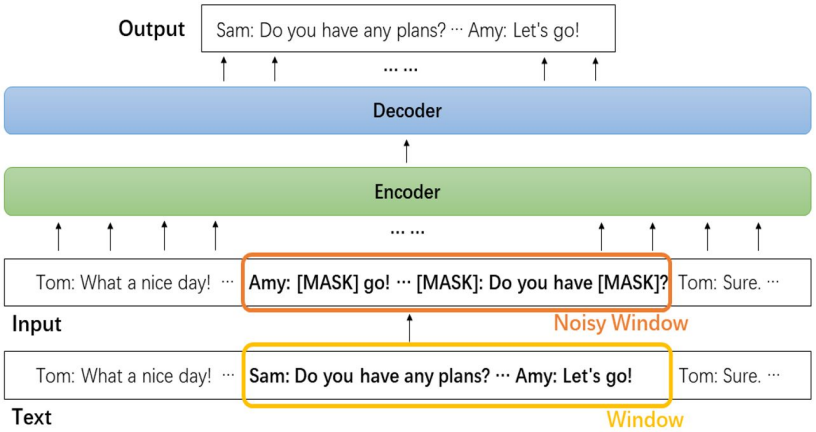
- Modify masking and noising methods to focus on turn-based transcript structure, e.g.,
  - Mask the whole turn
  - Randomly shuffle neighboring turns
  - Mask speakers

- Goal: recover the original turn & speaker info



Pretraining: DialogLM

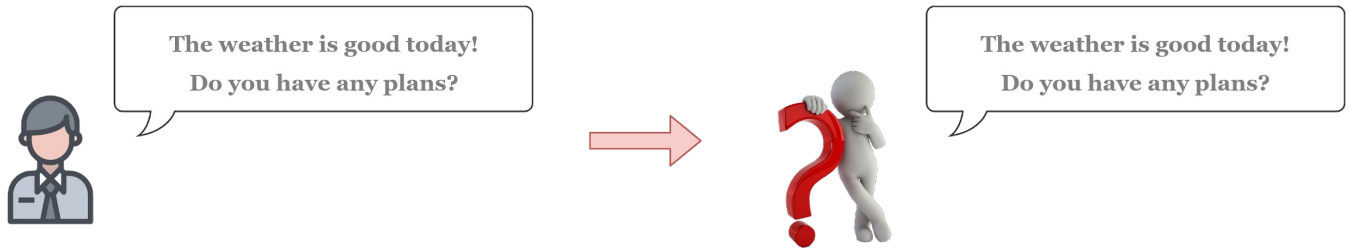
- Window-based Denoising
  - A window consists of consecutive turns
  - All noises are applied to this window
  - Input: the whole dialogue with the noisy window
  - Output: the denoised window



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

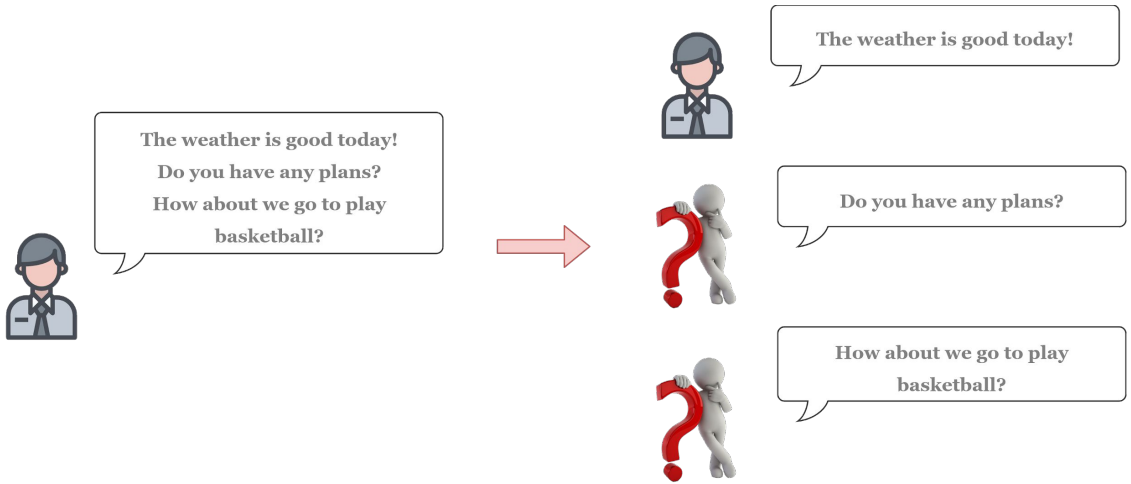
Pretraining: DialogLM

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



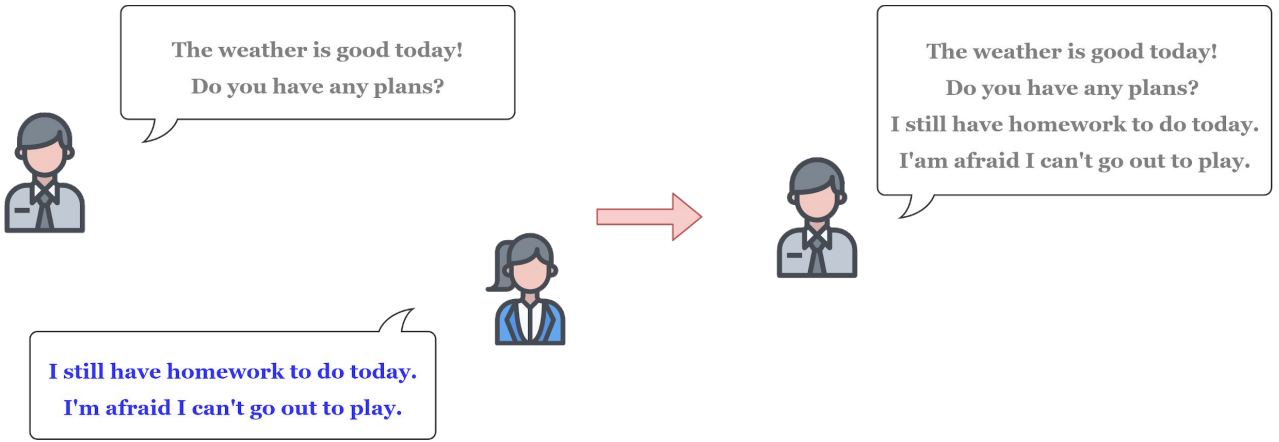
Pretraining: DialogLM

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



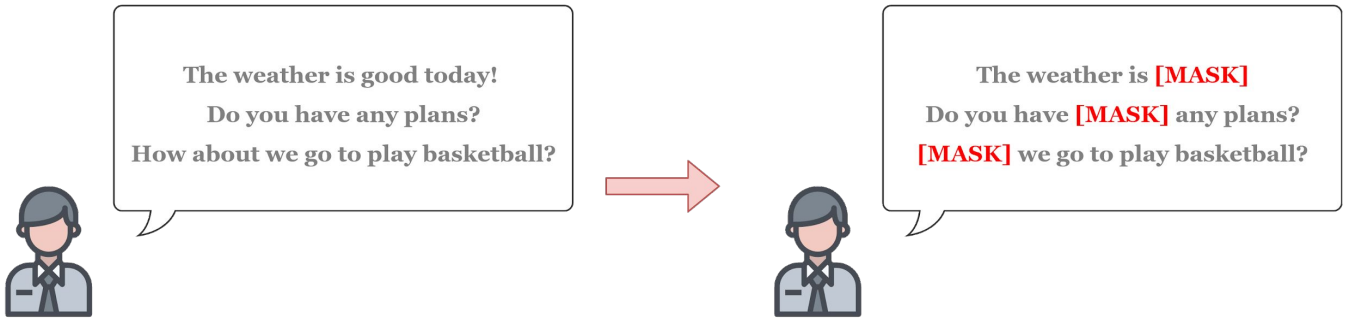
Pretraining: DialogLM

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



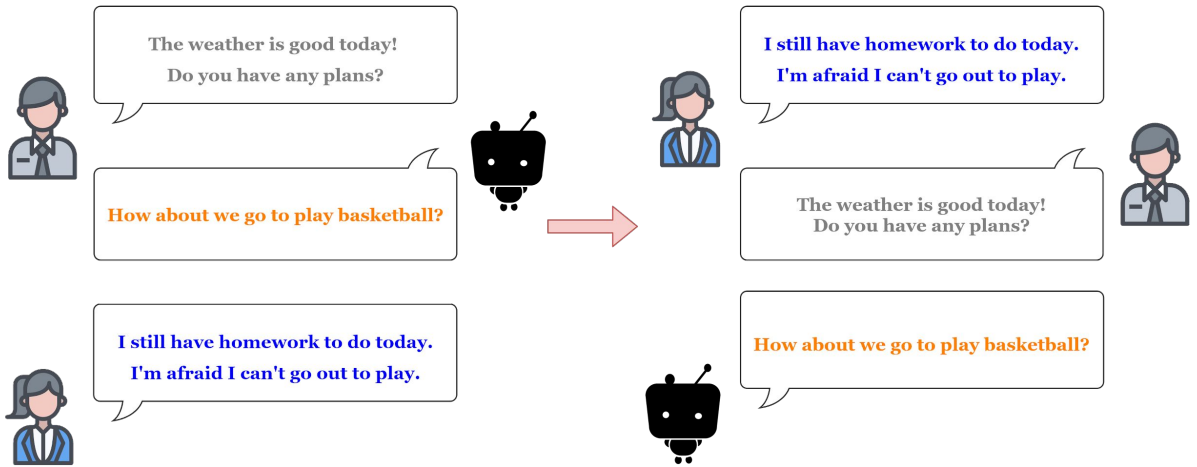
Pretraining: DialogLM

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



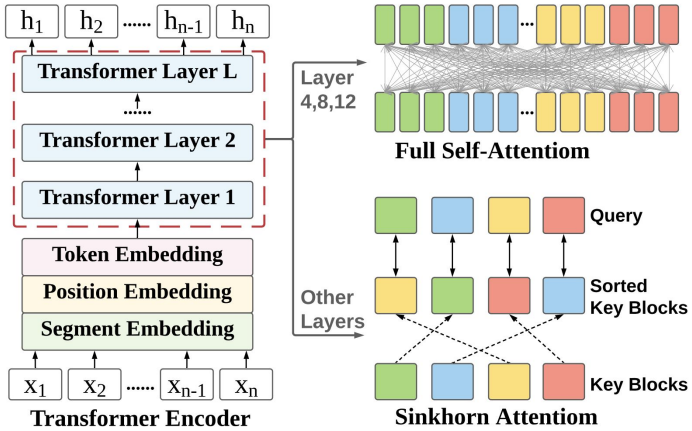
Pretraining: DialogLM

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



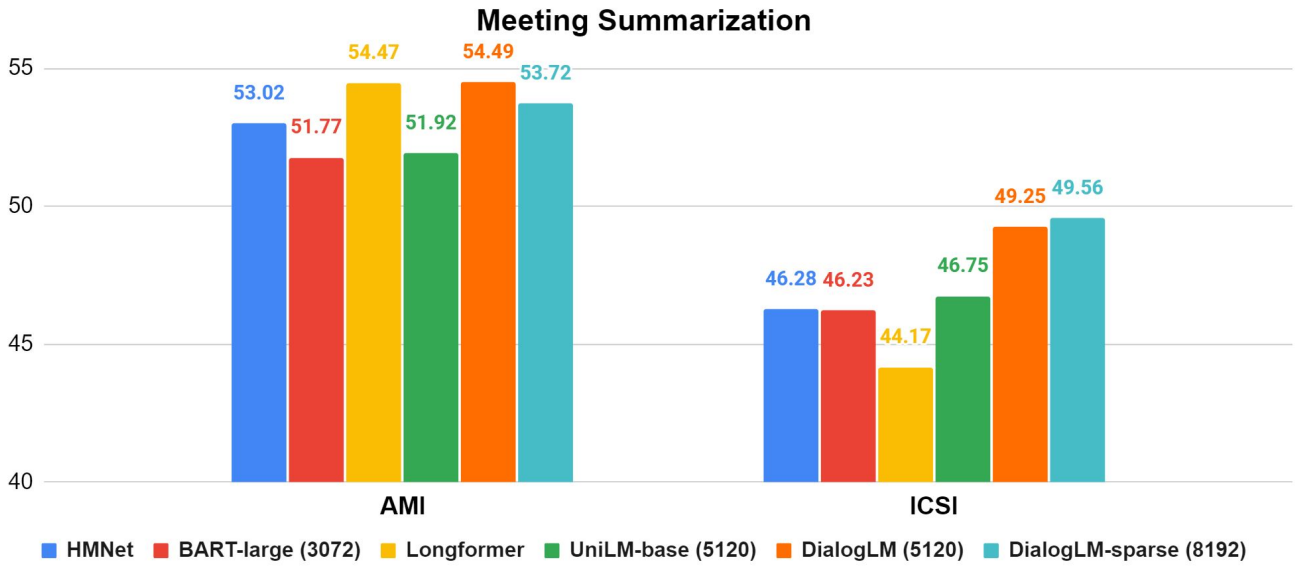
Pretraining: DialogLM

- 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



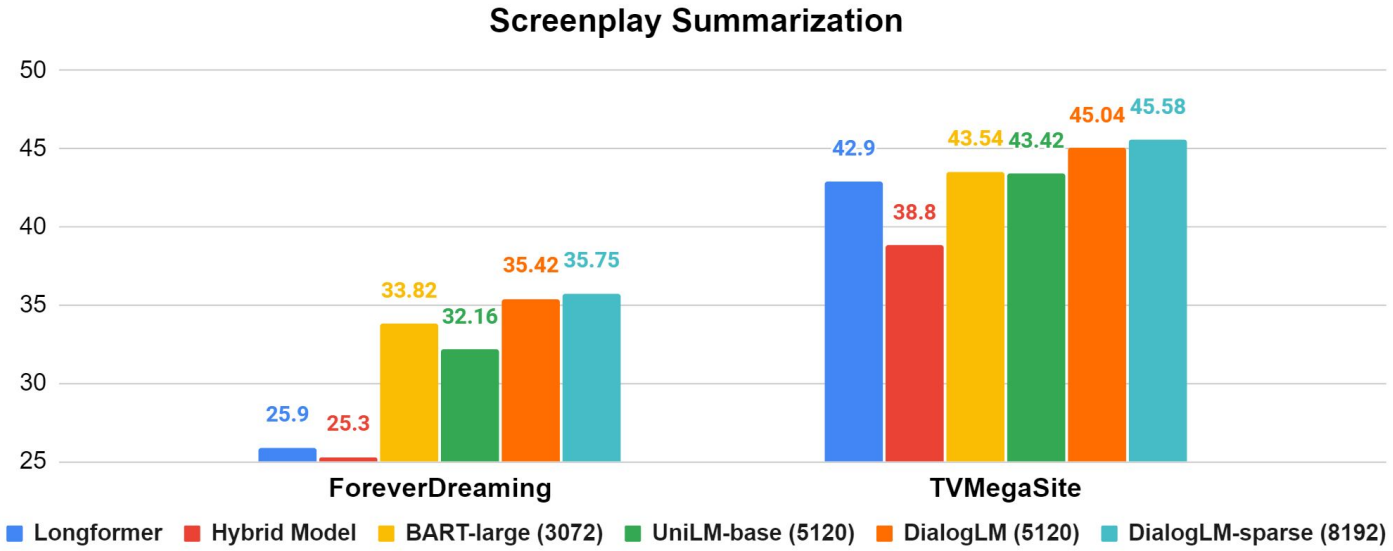
Pretraining: DialogLM

- Meeting summarization tasks



Pretraining: DialogLM

- Screenplay summarization tasks



Pretraining: HMNet

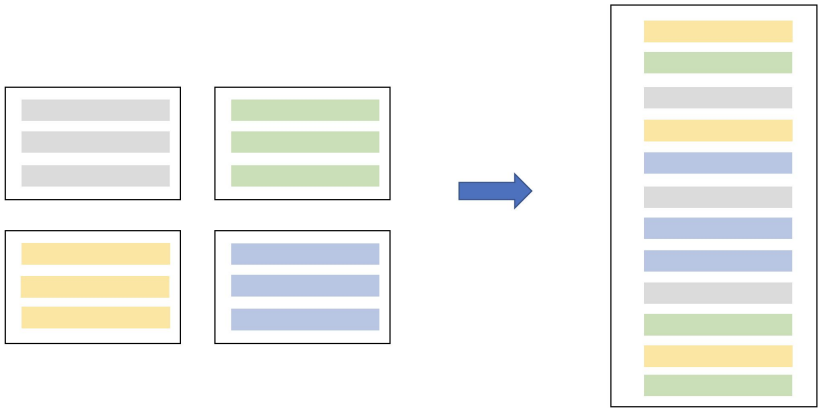
- When even pretraining data for target domain is scarce, cross-domain pretraining can help

- 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

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

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])

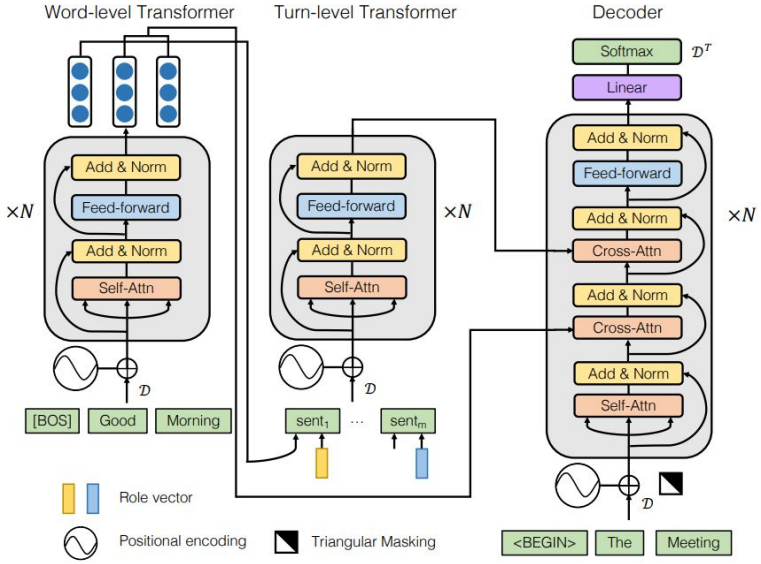
Long Dialogue

- 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

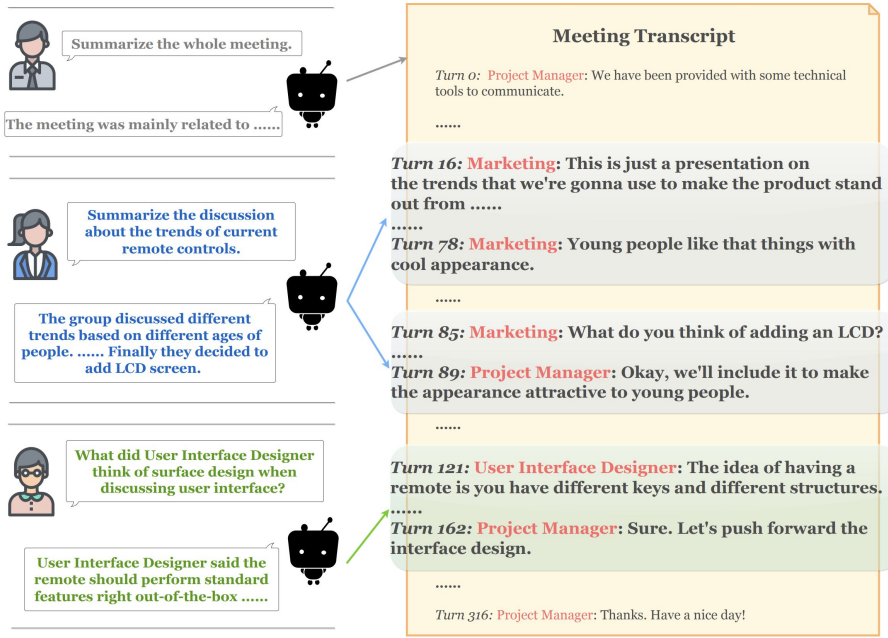
Dataset	AMI	
Source	HMNet	UNS
Readability	**4.17** (.38)	2.19 (.57)
Relevance	**4.08** (.45)	2.47 (.67)
Dataset	ICSI	
Source	HMNet	UNS
Readability	**4.24** (.20)	2.08 (.20)
Relevance	**4.02** (.55)	1.75 (.61)

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

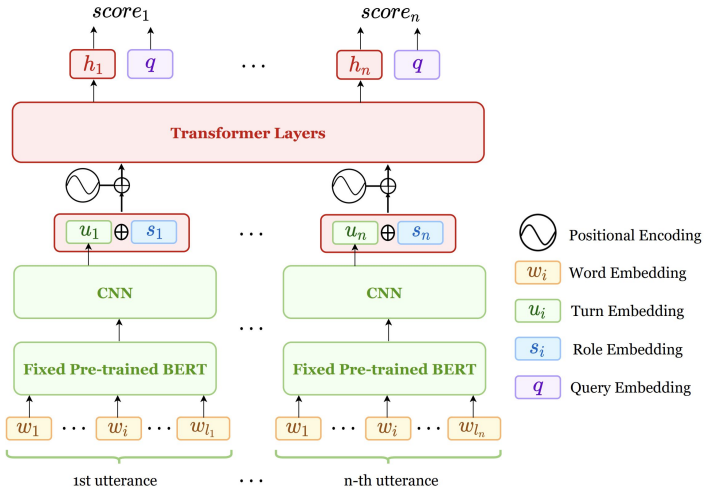
- There are multiple aspects of a meeting
- Challenging to cover all content in one single summary
- Solution: Query-based summarization
- Given the transcript and a query, focus on parts of the meeting related to the query and generate the summary answering that question



[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**

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**

Dataset: QMSum

Hierarchical ranking-based locator is the best

*: use retrieved text spans by locator

†: use gold text spans

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

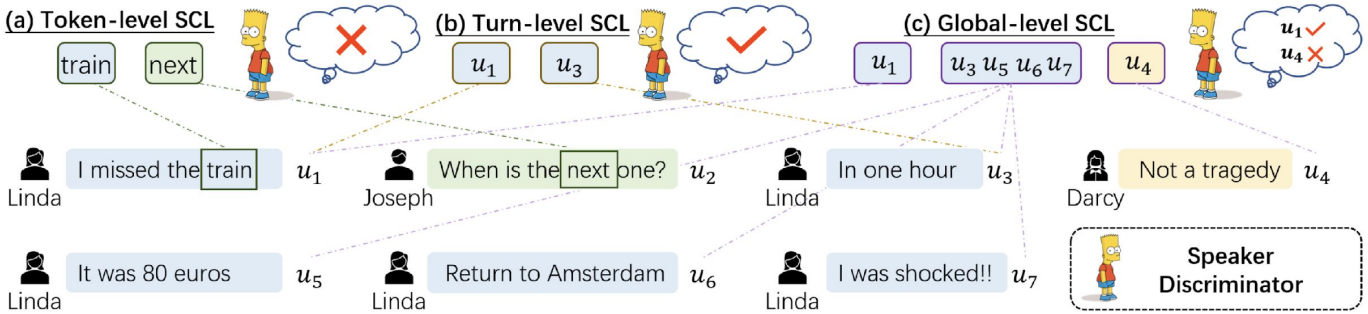
- Experiments show that existing dialogue models can hardly differentiate between different speakers
  - Aggregated BART embedding of K sampled tokens from utterance A and B
  - MLP classifier to judge whether A and B are from the same speaker
  - Random guess accuracy is **50%**
  - Classifier accuracy 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)
  - (**oi**, si) 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!
... ...

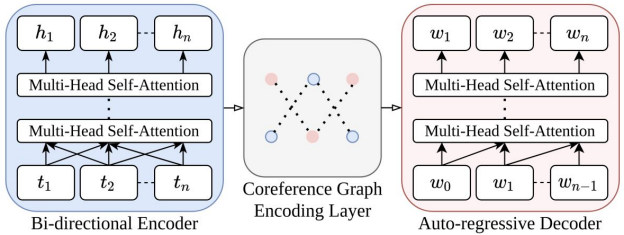
[6] *Coreference-Aware Dialogue Summarization*. Z. Liu et al. SIGDIAL 2021

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 in the coreference chain
- 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

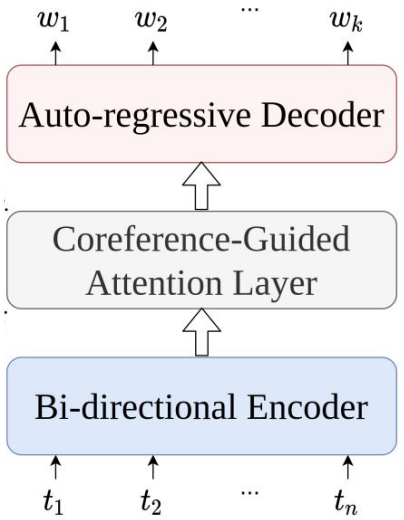


Multiple Participants: Coref

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

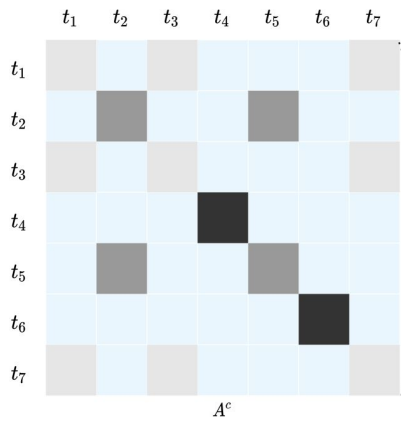
$$a_i = \sum_{j \in C^*} \frac{1}{|C^*|} h_j \text{ for } i \in C^*$$

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



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
Dataset: SAMSum

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

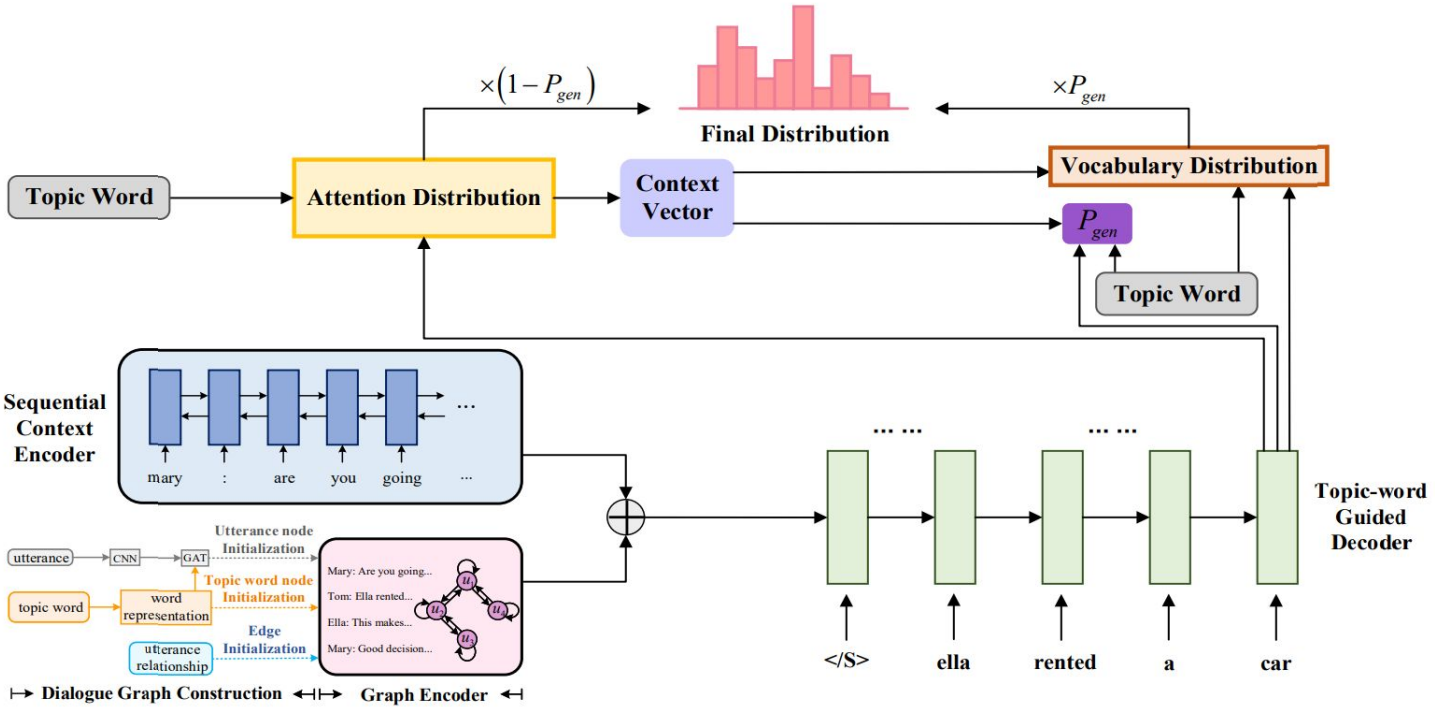
Knowledge Integration: TGDGA

- 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]	Contrastive 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

LLM as Dialogue Summarizer

Large Language Model (LLM)

- Models: GPT-3.5, ChatGPT, GPT-4, LLaMA, Alpaca, Vicuna, …
- Strong general language understanding and generation capabilities
- Follow various user instructions
- Can be directly used for text and dialogue summarization

Usage

- Instruction and input as prompt for zero-shot or few-shot learning
- Example:
  - *Please summarize the following dialogue in several sentences.*
  - *DIALOGUE: [Dialogue]*

LLM as Dialogue Summarizer

LLM as Dialogue Summarizer

- LLM is still behind fine-tuned models in terms of automatic metrics
- Output length can significantly impact ROUGE

Table 9: ROUGE scores of different models on the summarization dataset: SAMSum. We compare zero-shot ChatGPT with GPT-3.5 (*Zero-Shot*), BART-large (*Fine-Tuned*) [Lewis et al., 2019], and CODA (*fine-tuned*) [Chen and Yang, 2021].

Model	*zero-shot*		*fine-tuned*	
	ChatGPT	GPT-3.5	BART	CODA
ROUGE-1	41.0	**44.0**	49.1	**50.1**
ROUGE-2	17.4	**18.5**	24.3	**24.6**
ROUGE-L	32.2	**34.7**	45.8	**46.9**

Beatrice: I am in town, shopping. They have nice scarfs in the shop next to the church. Do you want one? Leo: No, thanks Beatrice: But you don't have a scarf. Leo: Because I don't need it. Beatrice: Last winter you had a cold all the time. A scarf could help. Leo: I don't like them. Beatrice: Actually, I don't care. You will get a scarf. Leo: How understanding of you! Beatrice: You were complaining the whole winter that you're going to die. I've had enough. Leo: Eh.

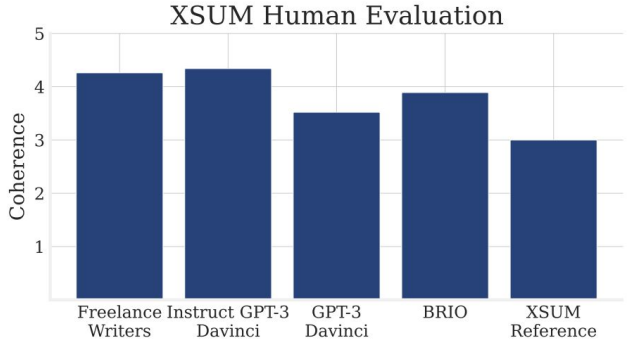
TL;DR: Leo is shopping in town with Beatrice and a friend. The friend suggests buying a scarf, but Leo does not want one because he does not like them. Beatrice insists on buying him one despite his reluctance, citing that he was constantly cold last winter and complained about it.

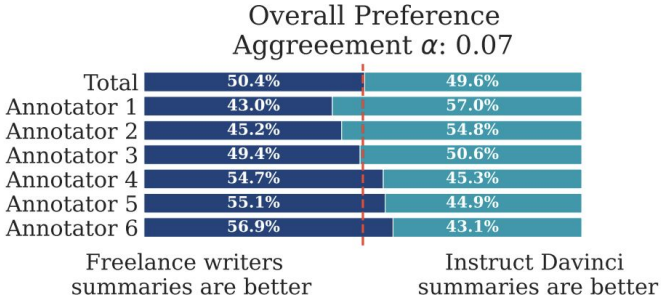
[12] *Is ChatGPT a General-Purpose Natural Language Processing Task Solver?*
C. Qin et al. arXiv 2302.06476

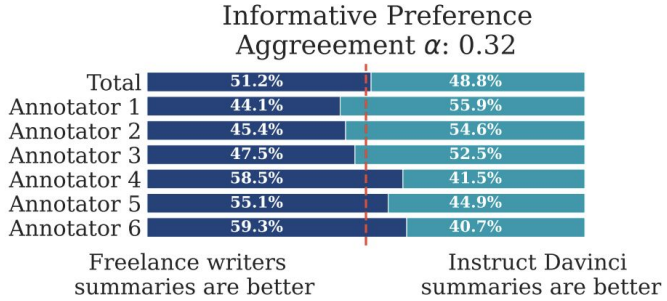
Human Evaluation

LLM as a News Summarizer by Human Evaluation

- On news summarization, GPT-3.5's summary (Instruct GPT-3) is already on par with expert human summary, **judged by humans**
- " … we may have reached the limit of single document news summarization"







[13] *Benchmarking Large Language Models for News Summarization*
T. Zhang et al. arXiv 2301.13848

Problems and Solutions of LLM

Existing Problems with LLM as a Summarizer

- Hallucination is a serious problem
  - Intrinsic hallucination: distorts information in the dialogue
  - Extrinsic hallucination: add information absent from the dialogue
- High cost, data privacy
- Domain adaptation

Potential Solutions

- Hallucination detection and reduction via prompt engineering, post-processing, and model tuning (for open source LLM)
- Federated system to ensemble LLM and smaller models
- On-premise training and deployment, especially for open-source LLMs

Outline

1. Introduction to Conversation Summarization
2. Conversation Structures and Evaluation
3. Pre-training and Models
4. Conclusion and Future Directions

Conclusion

- Conversations are pervasive, so summarization is important for efficient information digestion

Conclusion

- Conversations are pervasive, so summarization is important for efficient information digestion
- Conversation summarization (CS) is related to text summarization, but has its own unique challenges

Conclusion

- Conversations are pervasive, so summarization is important for efficient information digestion
- Conversation summarization (CS) is related to text summarization, but has its own unique challenges
- The field of CS is booming, with more methods and datasets

Conclusion

- Conversations are pervasive, so summarization is important for efficient information digestion
- Conversation summarization (CS) is related to text summarization, but has its own unique challenges
- The field of CS is booming, with more methods and datasets
- CS benefits from LLM like other applications

Future Directions

- Multi-modal Input
  - textual transcripts
  - prosodic audios
  - visual videos

https://www.skillcast.com/blog/best-practices-productive-meetings

Future Directions

- Multi-modal Input
- Multi-domain Summarization
  - Email
  - Debate
  - Meeting
  - Interview
  - ...

Future Directions

- Multi-modal Input
- Multi-domain Summarization
- Multilingual Input

Future Directions

- Multi-modal Input
- Multi-domain Summarization
- Multi-lingual Input
- Privacy
- Efficiency

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

[12] Is ChatGPT a General-Purpose Natural Language Processing Task Solver? C. Qin et al. arXiv 2302.06476

[13] Benchmarking Large Language Models for News Summarization. T. Zhang et al. arXiv 2301.13848

Where can I find more information?

Talk slides, reading materials, schedule:
https://github.com/zcgzcgzcg1/EACL2023_Tutorial_Dialogue_Summarization

Tutorial Proposal
https://aclanthology.org/2023.eacl-tutorials.3.pdf