



AACL 2023 Tutorial

Learning **WHO** Saying **WHAT** to **WHOM** in Multi-Party Conversations

Jia-Chen Gu¹, Zhuosheng Zhang², Zhen-Hua Ling¹

¹University of Science and Technology of China

²Shanghai Jiao Tong University

November 1, 2023

Presenters



Jia-Chen Gu
Postdoc@USTC



Zhuosheng Zhang
Assit. Prof@SJTU



Zhen-Hua Ling
Prof@USTC

Participation and Q&A

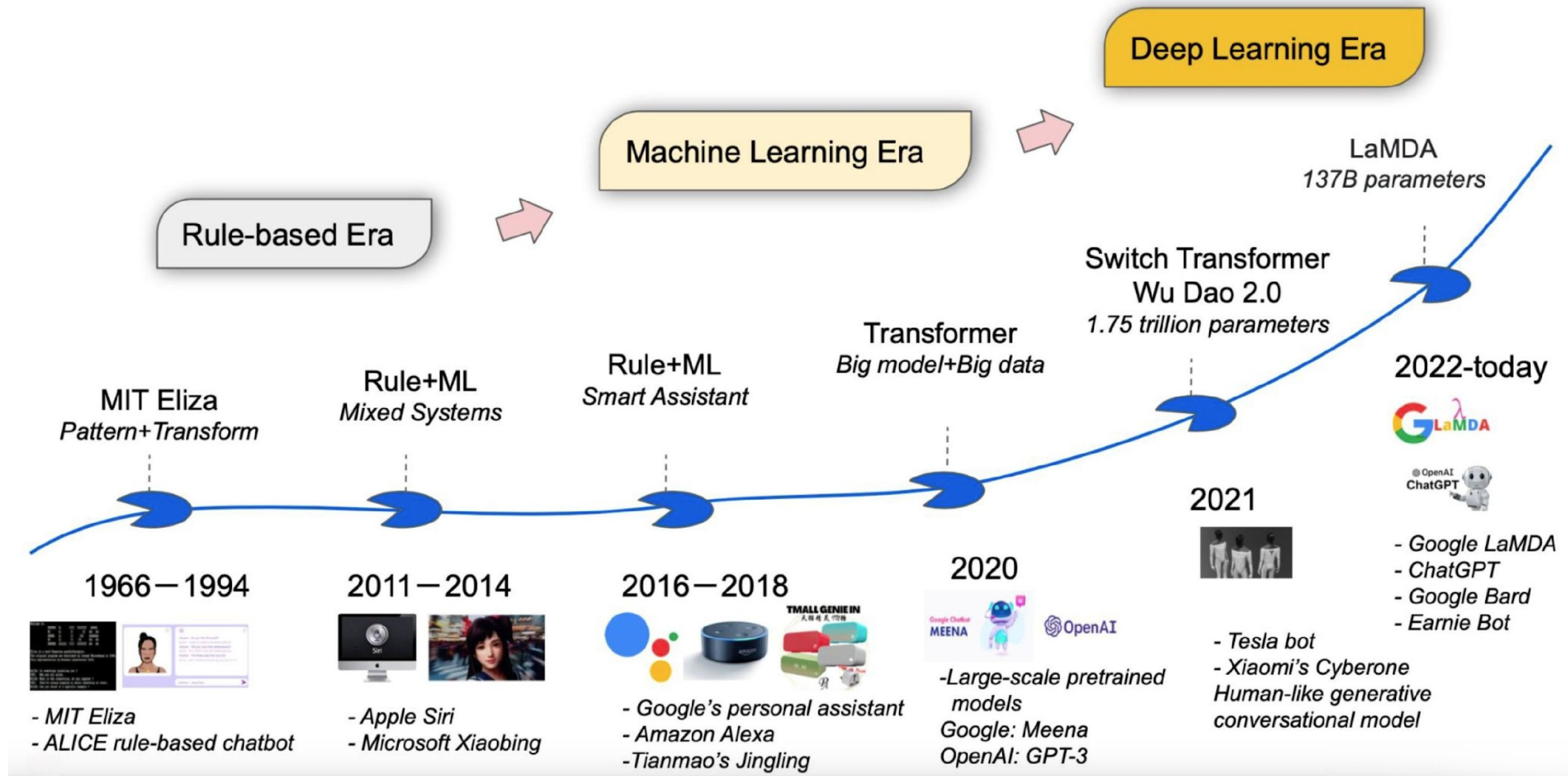
- All tutorial slides are available at:

http://home.ustc.edu.cn/~gujc/presentations/2023_AACL_MPC.pdf



- 10-minute Q&A session after each half

History of Conversational AI



Two-Party Conversations



One-on-one chat
between 2 interlocutors

Two-Party VS. Multi-Party Conversations



One-on-one chat
between 2 interlocutors



Group chat
involving 3+ interlocutors

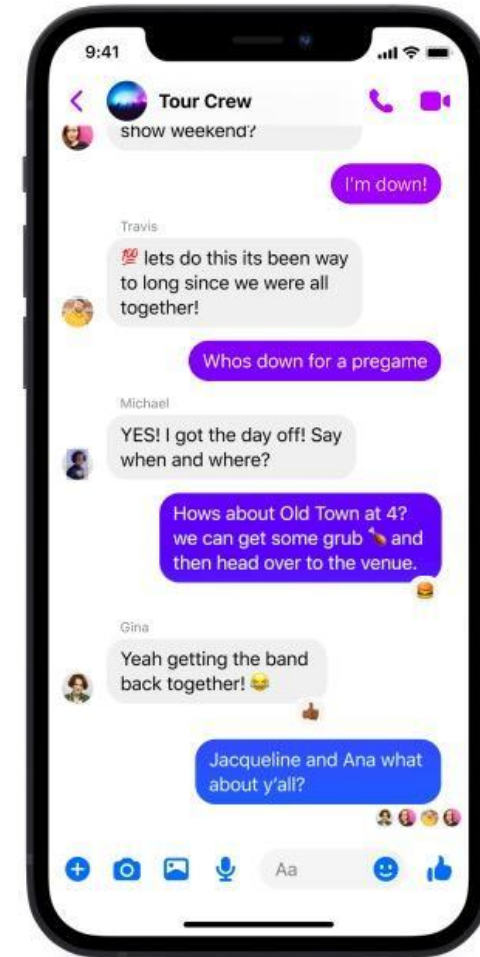
Two-Party VS. Multi-Party Conversations



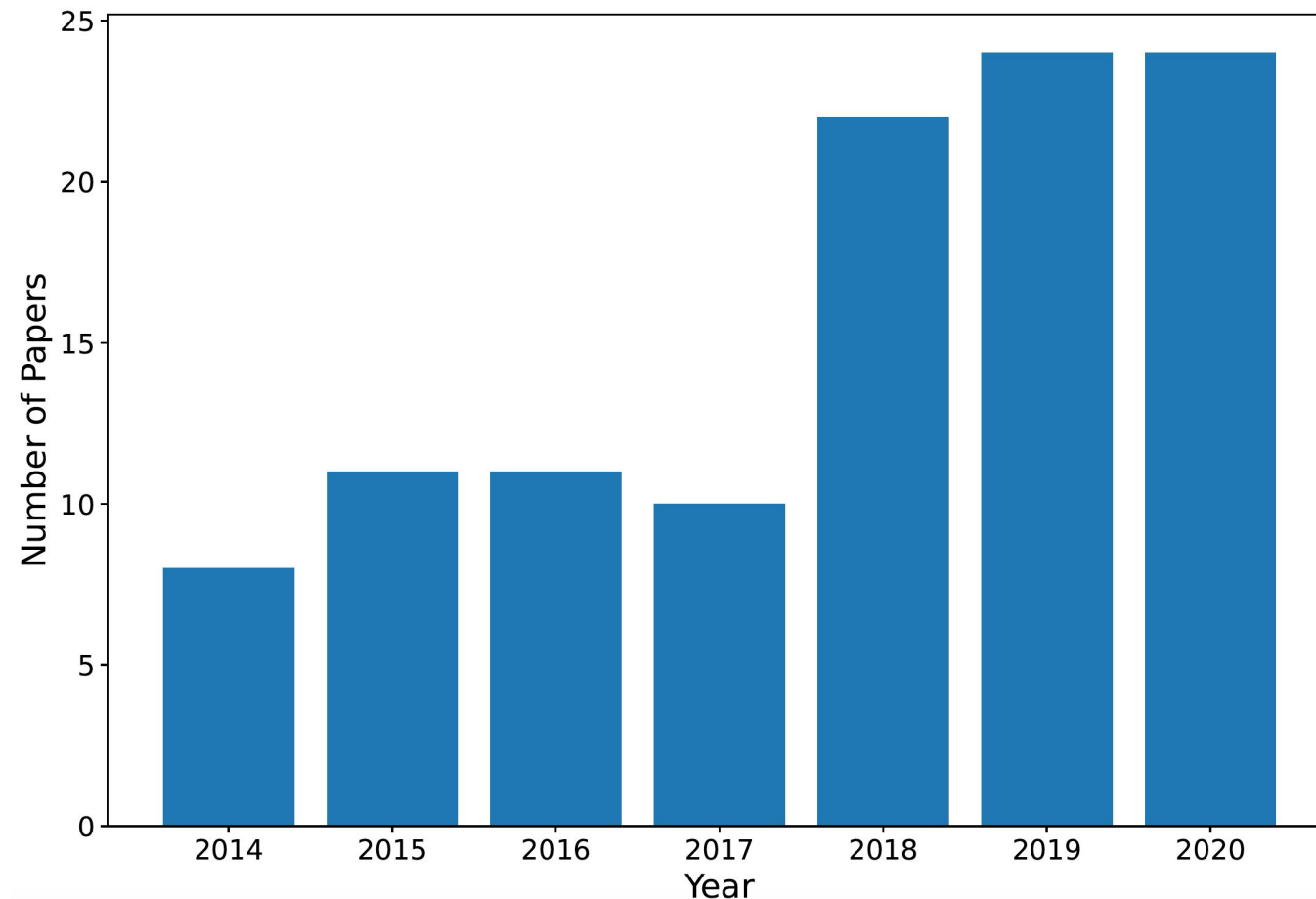
One-on-one chat
between 2 interlocutors

Group chats appear
frequently in daily life!

Group chat
involving 3+ interlocutors



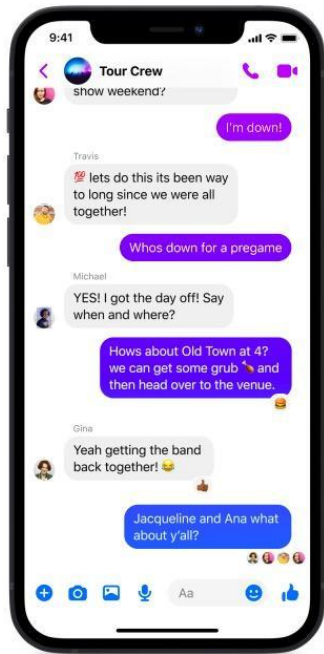
Research Trend on Multi-Party Conversation



Include the **keywords** **multi-party conversation** and its variations, considering papers that appeared at **CL and AI venues**

Why Multi-Party Conversations (MPC)?

Many scenarios involve MPC and require capabilities beyond two-party conversations, e.g., turn-taking, discourse parsing and disentanglement



Group Chat

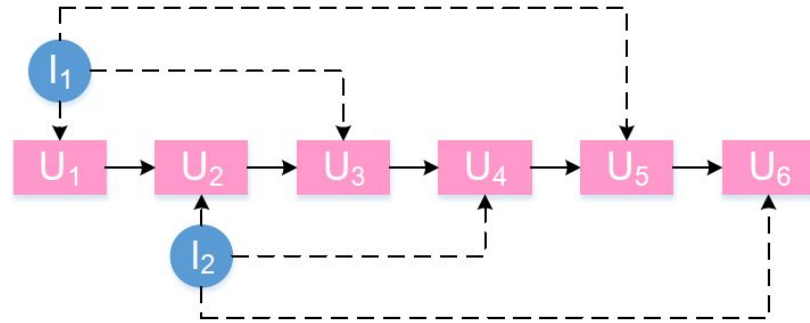


Meeting



Agent Simulacra

Graphical MPC is Complicated



Utterances in a **two-party** conversation are posted one by one between two interlocutors, constituting a **sequential** information flow

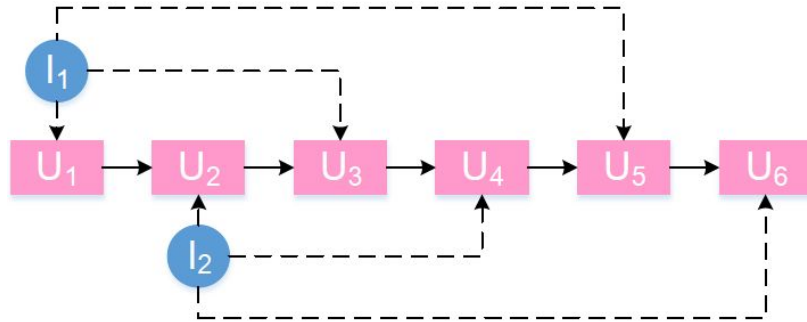


: Interlocutors

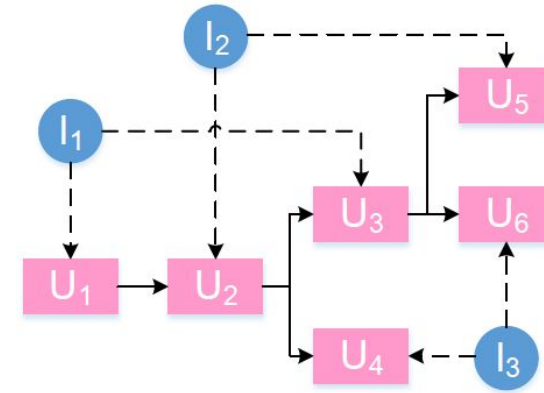


: Utterances

Graphical MPC is Complicated



Utterances in a **two-party** conversation are posted one by one between two interlocutors, constituting a **sequential** information flow



Utterances in a **multi-party** conversation can be spoken by anyone and address anyone else, constituting a **graphical** information flow



Challenges (1): WHO Speaks

Model the coordination strategies that speakers adopt to **acquire or give up the floor**, so that an ongoing conversation can go on smoothly (Hawes et al., 2009; Pinhanez et al., 2018; de Bayser et al., 2019)

Speaker	Addressee	Utterance
User 1	-	I have a problem when I install ...
Agent	User 1	Did you set initial params?
User 2	User 1	Show the error message, and ...
User 1	Agent	How?
User 1	User 2	OK, just a moment!
[Who speak?]		

Should the agent take
the floor to speak or
not?

Challenges (2): Address WHOM

Understand conversation semantics for the behavior whereby interlocutors **indicate to whom they are speaking** (Ouchi and Tsuboi, 2016; Le et al., 2019; Gu et al., 2021; Zhu et al., 2023)

Speaker	Addressee	Utterance
User 1	-	I have a problem when I install ...
Agent	User 1	Did you set initial params?
User 2	User 1	Show the error message, and ...
User 1	Agent	How?
User 1	User 2	OK, just a moment!
Agent	[To whom?]	

User 1?

or

User 2?

Challenges (3): Say WHAT

Return an appropriate response which follows the conversation **semantics**, **structures** and **topic transitions** (Zhang et al., 2018; Wu et al., 2020; Wang et al., 2020; Gu et al., 2022; Li et al., 2023)

Speaker	Addressee	Utterance
User 1	-	I have a problem when I install ...
Agent	User 1	Did you set initial params?
User 2	User 1	Show the error message, and ...
User 1	Agent	How?
User 1	User 2	OK, just a moment!
Agent	User 1	[Say what?]

See this URL: <http://xxx>
or
It's already in OS

Challenges (3): Say WHAT

Return an appropriate response which follows the conversation **semantics**, **structures** and **topic transitions** (Zhang et al., 2018; Wu et al., 2020; Wang et al., 2020; Gu et al., 2022; Li et al., 2023)

Speaker	Addressee	Utterance
User 1	-	I have a problem when I install ...
Agent	User 1	Did you set initial params?
User 2	User 1	Show the error message, and ...
User 1	Agent	How?
User 1	User 2	OK, just a moment!
Agent	User 1	See this URL: http://xxx

Goals of Tutorial

- We will cover a number of key developments on multi-party conversations (mostly 2018–2023)
 - ✓ **WHO speaks**
 - ✓ **address WHOM**
 - ✓ **say WHAT**
- This tutorial is **cutting-edge**, and we are still far from understanding how to best develop multi-party conversational AI:
 - ✓ **Taxonomies of existing research and key insights**
 - ✓ **Our perspectives on the current challenges & open problems**



Schedule

Time	Section	Presenter
14:00—14:15	Section 1: Introduction	Zhen-Hua Ling
14:15—14:20	Section 2: Datasets	Jia-Chen Gu
14:20—14:40	Section 3: Speaker Modeling	Jia-Chen Gu
14:40—15:20	Section 4: Addressee Modeling	Jia-Chen Gu
15:20—15:30	Q & A Session I	All
30 mins Coffee Break		
16:00—16:30	Section 4: Addressee Modeling (cont'd)	Zhuosheng Zhang
16:30—17:05	Section 5: Response Modeling	Zhuosheng Zhang
17:05—17:20	Section 6: Challenges & Opportunities	Zhen-Hua Ling
17:20—17:30	Q & A Session II	All

Section 2: Datasets

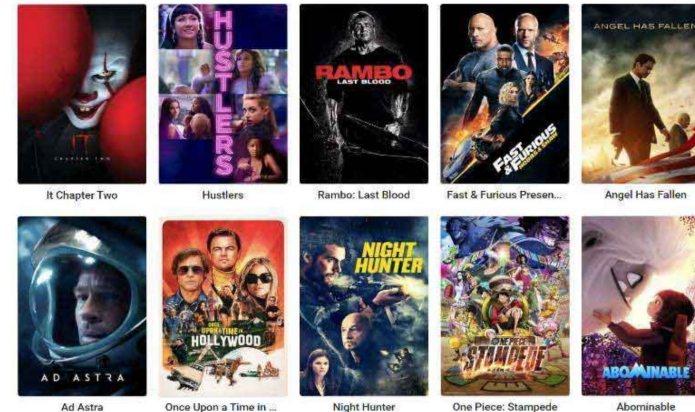
Datasets

- **Written** corpora: **online forums**, such as Ubuntu, Reddit ...



Datasets

- **Written** corpora: **online forums**, such as Ubuntu, Reddit ...



- **Spoken** corpora:
 - ✓ **scripted** refers to planned dialogue, such as TV and movie scripts
 - ✓ **Unscripted** refers to spontaneous and unplanned dialogues, such as meeting

Written Corpora

Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
NPS Chat Corpus	Informal chat	15	100M			×	part-of-speech tagging, dialogue act recognition
Ubuntu Dialogue Corpus	Ubuntu OS Chatroom	930K	100M	-	-	×	speaker identification, discourse parsing, machine comprehension, response selection
Ubuntu Chat Corpus	Ubuntu OS Chatroom	10655	2B	-	-	×	language learning, POS tagging
Molweni	Ubuntu OS Chatroom	10K	24K	200 hrs	3.5 per dialogue	×	machine reading comprehension, discourse parsing
MPC Corpus	Informal chatroom	14	58K	-	5 per session	×	turn-taking, speaker identification, detecting influence & leadership, group behavior
Settlers of Catan	Informal, game-playing	21	-	-	2-6 players	×	modeling bargaining, negotiation, trading dialogue, risk-management in dialogue, action identification
Cards Corpus	Informal, game-playing	1266	282K	-	-	×	goal-driven dialogue, event knowledge based questioning
Reddit Corpus	Informal forum	84979	76M-414M*	-	521K	Maybe	discourse, cyberbully detection, exploring incel language
Reddit Domestic Abuse Corpus	Abusive forum	21333	19M-303M	-		×	language biases, detecting harassment
Internet Argument Corpus	Political forum	11000	73M	-	-	×	summarization, rhetoric and sarcasm, stance detection
Agreement in Wikipedia Talk Pages	Informal	822	110K	-	-	×	linguistic tracing of manipulations, dialog act recognition, social act recognition, conflict detection, speaker identification
Agreement by Create Debaters	Informal	10000	1.4M	-	-	×	constructive disagreement, sarcasm, rumor classification, stance identification
Twitter Corpus	Informal microblog	1.3M	125M	-	-	×	dialogue act recognition, author and topic identification, event discovery
UseNet Corpus	Informal microblog	47860	7B	-	-	×	modeling and analyzing text written on mobile devices

Spoken Corpora

UNSCRIPTED INFORMAL	Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
	Aggregated from various sources							
	British National Corpus (BNC)	Informal	854	10M	100 hrs*	23466	✓	word sense disambiguation, morphological & syntactic analysis
	CANCODE	Informal	-	5M	550 hrs*	-	×	language learning, POS tagging
	Collected in specialized environments							
	D64 Corpus	Natural	2	70K*	8 hrs	5	✓	involvement detection, studying silence and overlap in conversation
	COSINE	Natural	10	160K	42 hrs	3.69 per session	✓	recognition of speech and speakers in noisy environments
	IDIAP Wolf Corpus	Game	15	60K*	7 hrs	8-12 groups	✓	group performance in task-based interaction, implicit communication
	TEAMS corpus	Game	116K	3M	47 hrs	3-4/ game	✓	entrainment, speaker transitions, personality identification & team dynamics
	Transcribed from pre-recorded media							
	COLT corpus	Natural	100	500K	55 hrs	31	×	teenage talk trends
	CRD3	Game	159	5M	-	72	✓	character-action interactions in role playing games
SCRIPTED FORMAL	Transcribed from pre-recorded media							
	Intelligence Squared Debates	Debates, predecided	108	1.8M	200 hrs*	3-5 per debate	✓	predictive models of debates, discourse modeling
	CSPAIE	Politics, education	200	2M	220 hrs*	400+	×	speech style and gender distinctions, speech variation between written and spoken corpora
	CED (1560-1760)	Movies, formal	-	1.2 M	-	-	×	early English language variations and changes over time
	MediaSum	Interview	463K	720M	-	6.5 per dialogue	✓	dialogue summarization
	INTERVIEW corpus	Interview	105K	126.7M	10K	184K	✓	follow-up question generation
	Canal9	Political Debates	70 debates	-	43 hrs	5 per debate	✓	speaker identification, turn-taking, conflict detection
	Transcribed from pre-recorded media							
	Movie-DiC	Movie dialogues	132K	6M	-	1-7 per dialogue	×	turn taking, speaker identification, emotional dialogue generation
	Cornell Movie Dialogue Corpus	Movie dialogues	220K	9M	-	9035	×	
SCRIPTED SPOKEN	Film scripts online series	Movie scripts	263K	16M	1500 scripts	2-6 per script*	×	(information unavailable)
	OpenSubtitles	Movie subtitles	337M	2.5G	-	2-6 per script*	×	turn taking, speaker identification, emotional dialogue generation
	SubTle corpus	Movie subtitles	3.35M	20M	6184 movies	2-6 per script*	×	
	Character Style from Film Corpus	Movie subtitles	151K	9.6M	862 movies	2-6 per script*	×	
	American Soap Opera Corpus	TV dialogues	1.2M	100M	-	10-12 per script	×	
	TVD corpus	TV dialogues	10K	600K	-	2-6 per script	✓	turn taking, speaker identification, emotional dialogue generation
	MELD	TV dialogues	1400	109K	13.6 hrs*	400	✓	
	Serial Speakers	TV dialogues	106K	682K	130 hrs	6 per script*	✓	
	MEISD	TV dialogues	1000	50K unique	22 hrs	4072	✓	

UNSCRIPTED INFORMAL	Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
	Aggregated from various sources							
	British National Corpus (BNC)	Informal	854	10M	100 hrs*	23466	✓	word sense disambiguation, morphological & syntactic analysis
	CANCODE	Informal	-	5M	550 hrs*	-	×	language learning, POS tagging
	Collected in specialized environments							
	D64 Corpus	Natural	2	70K*	8 hrs	5	✓	involvement detection, studying silence and overlap in conversation
	COSINE	Natural	10	160K	42 hrs	3.69 per session	✓	recognition of speech and speakers in noisy environments
	IDIAP Wolf Corpus	Game	15	60K*	7 hrs	8-12 groups	✓	group performance in task-based interaction, implicit communication
	TEAMS corpus	Game	116K	3M	47 hrs	3-4/ game	✓	entrainment, speaker transitions, personality identification & team dynamics
	Transcribed from pre-recorded media							
	COLT corpus	Natural	100	500K	55 hrs	31	×	teenage talk trends
	CRD3	Game	159	5M	-	72	✓	character-action interactions in role playing games
SCRIPTED FORMAL	Transcribed from pre-recorded media							
	Intelligence Squared Debates	Debates, predecided	108	1.8M	200 hrs*	3-5 per debate	✓	predictive models of debates, discourse modeling
	CSPAIE	Politics, education	200	2M	220 hrs*	400+	×	speech style and gender distinctions, speech variation between written and spoken corpora
	CED (1560-1760)	Movies, formal	-	1.2 M	-	-	×	early English language variations and changes over time
	MediaSum	Interview	463K	720M	-	6.5 per dialogue	✓	dialogue summarization
	INTERVIEW corpus	Interview	105K	126.7M	10K	184K	✓	follow-up question generation
	Canal9	Political Debates	70 debates	-	43 hrs	5 per debate	✓	speaker identification, turn-taking, conflict detection
	Transcribed from pre-recorded media							
	Movie-DiC	Movie dialogues	132K	6M	-	1-7 per dialogue	×	turn taking, speaker identification, emotional dialogue generation
	Cornell Movie Dialogue Corpus	Movie dialogues	220K	9M	-	9035	×	
SCRIPTED SPOKEN	Film scripts online series	Movie scripts	263K	16M	1500 scripts	2-6 per script*	×	(information unavailable)
	OpenSubtitles	Movie subtitles	337M	2.5G	-	2-6 per script*	×	turn taking, speaker identification, emotional dialogue generation
	SubTle corpus	Movie subtitles	3.35M	20M	6184 movies	2-6 per script*	×	
	Character Style from Film Corpus	Movie subtitles	151K	9.6M	862 movies	2-6 per script*	×	
	American Soap Opera Corpus	TV dialogues	1.2M	100M	-	10-12 per script	×	
	TVD corpus	TV dialogues	10K	600K	-	2-6 per script	✓	turn taking, speaker identification, emotional dialogue generation
	MELD	TV dialogues	1400	109K	13.6 hrs*	400	✓	
	Serial Speakers	TV dialogues	106K	682K	130 hrs	6 per script*	✓	
	MEISD	TV dialogues	1000	50K unique	22 hrs	4072	✓	

Ubuntu IRC

Text-based discussions on
posting specialized questions
and replying with solutions

- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command

Ubuntu IRC

Text-based discussions on posting specialized questions and replying with solutions

- conversation disentanglement (variation of information, one-to-one overlap, and exact match F_1 for evaluating clusters)

- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command



- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command

Ubuntu IRC

Text-based discussions on posting specialized questions and replying with solutions

- conversation disentanglement (variation of information, one-to-one overlap, and exact match F_1 for evaluating clusters)
- speaker identification (precision, accuracy)

- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command



- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command

Ubuntu IRC

Text-based discussions on posting specialized questions and replying with solutions

- conversation disentanglement (variation of information, one-to-one overlap, and exact match F_1 for evaluating clusters)
- speaker identification (precision, accuracy)
- addressee recognition (precision, accuracy)

- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...@...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...@...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command



- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...@...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...@...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command

Ubuntu IRC

Text-based discussions on posting specialized questions and replying with solutions

- conversation disentanglement (variation of information, one-to-one overlap, and exact match F_1 for evaluating clusters)
- speaker identification (precision, accuracy)
- addressee recognition (precision, accuracy)
- response generation (auto e.g., BLEU and ROUGE, and human evaluations e.g., relevance and informativeness)

- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command



- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- === E-bola [...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] delire: BurgerMann, i used that to convert 100 of photos in one command

MELD

- **Multi-modal** conversations from the **TV show *Friends*** (video, audio, text)
- **Emotion recognition** (anger, disgust, fear, joy, neutral, sadness, surprise)
- **Sentiment recognition** (positive, negative, neutral)

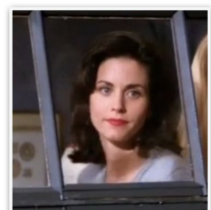
Metric: **F-score**



Utterance: "Become a drama critic!"

Emotion: Joy Sentiment: Positive

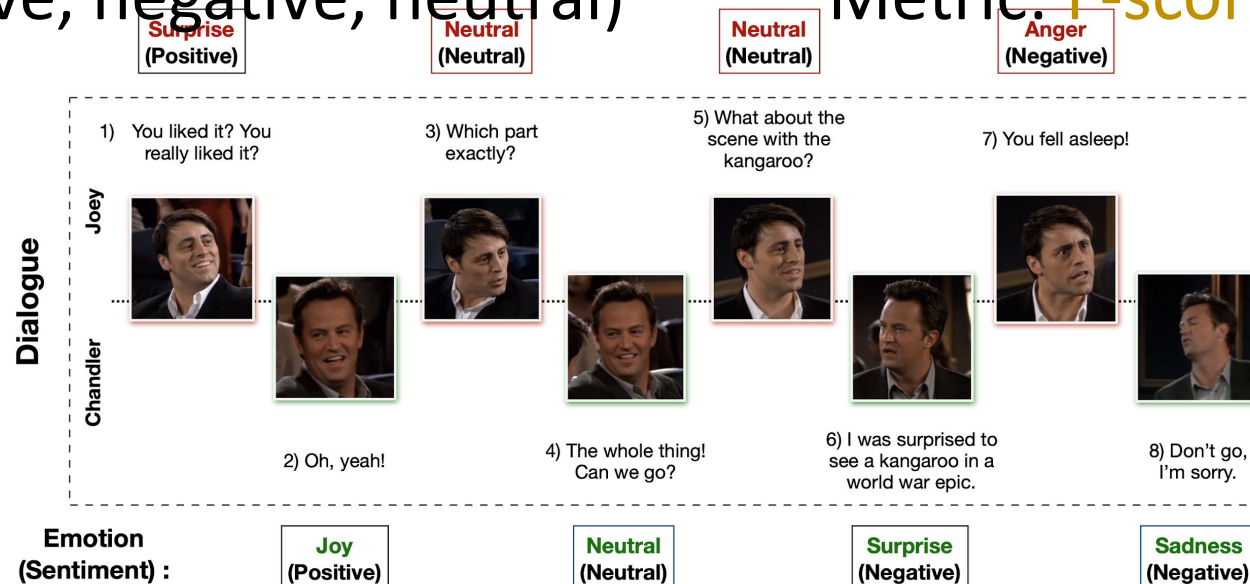
Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face



Utterance: "Great, now he is waving back"

Emotion: Disgust Sentiment: Negative

Text	Audio	Visual
Positive/Joy	Flat tone	Frown



Challenges of Data Collection

- **Informative metadata** to **track interlocutors** (speaker, addressee, observer) within and across (if possible) dialogues

Challenges of Data Collection

- **Informative metadata** to track interlocutors (speaker, addressee, observer) within and across (if possible) dialogues
- **Balanced Interlocutors** in terms of various demographic factors, so as to not preferentially model any specific type of language

Challenges of Data Collection

- **Informative metadata** to **track interlocutors** (speaker, addressee, observer) within and across (if possible) dialogues
- **Balanced Interlocutors** in terms of various **demographic factors**, so as to not preferentially model any specific type of language
- **High quality** text to **prevent noise from drowning out important signals**, particularly for data scraped from the web

Challenges of Data Collection

- **Informative metadata** to track interlocutors (speaker, addressee, observer) within and across (if possible) dialogues
- **Balanced Interlocutors** in terms of various demographic factors, so as to not preferentially model any specific type of language
- **High quality** text to prevent noise from drowning out important signals, particularly for data scraped from the web
- **Privacy protection** to avoid leaking personally identifiable information, especially for multi-modal corpora

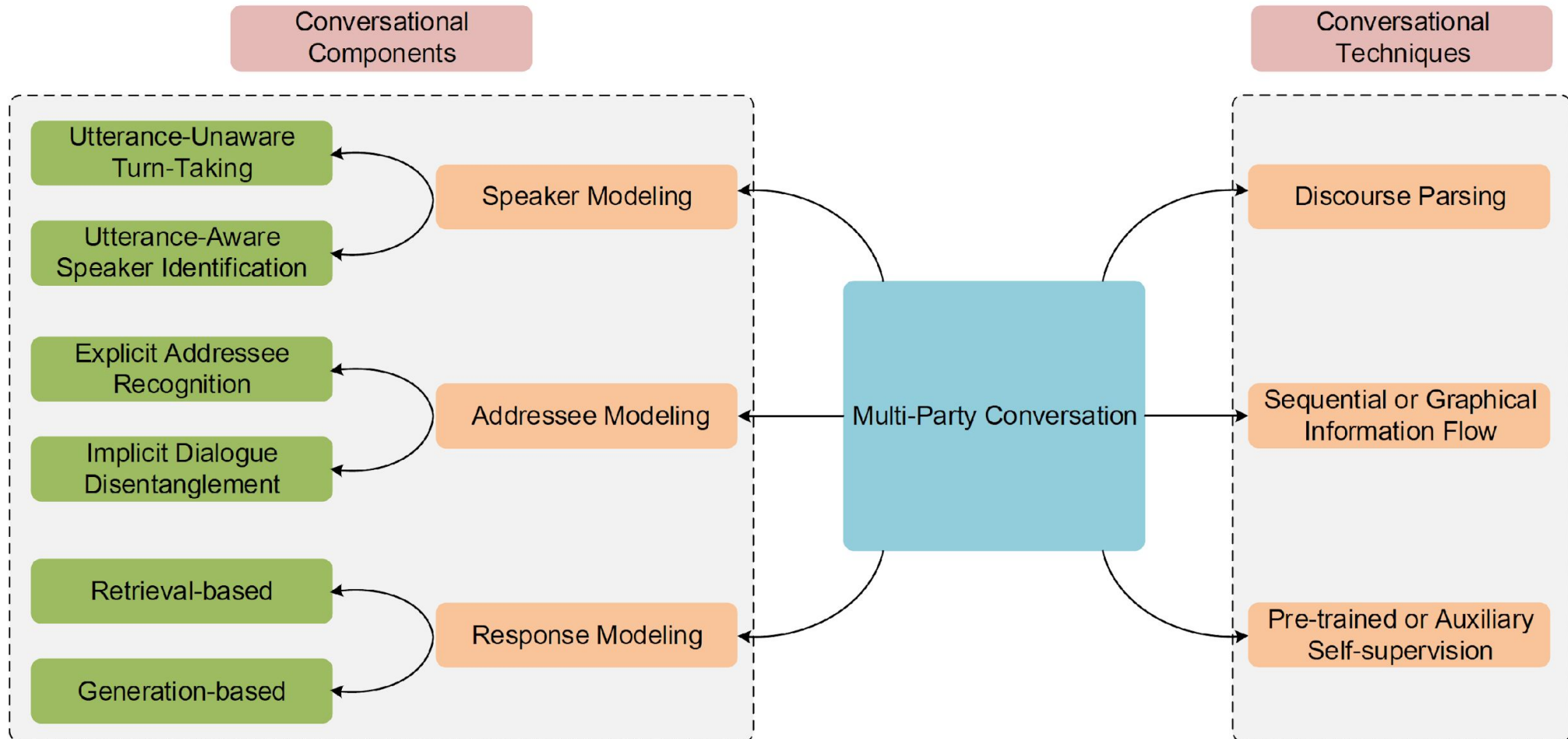
Challenges of Data Collection

- **Informative metadata** to track interlocutors (speaker, addressee, observer) within and across (if possible) dialogues
- **Balanced Interlocutors** in terms of various demographic factors, so as to not preferentially model any specific type of language
- **High quality** text to prevent noise from drowning out important signals, particularly for data scraped from the web
- **Privacy protection** to avoid leaking personally identifiable information, especially for multi-modal corpora

Refer detailed dataset taxonomy to

On the Need for Thoughtful Data Collection for Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods (SIGDIAL 2021)

Overview



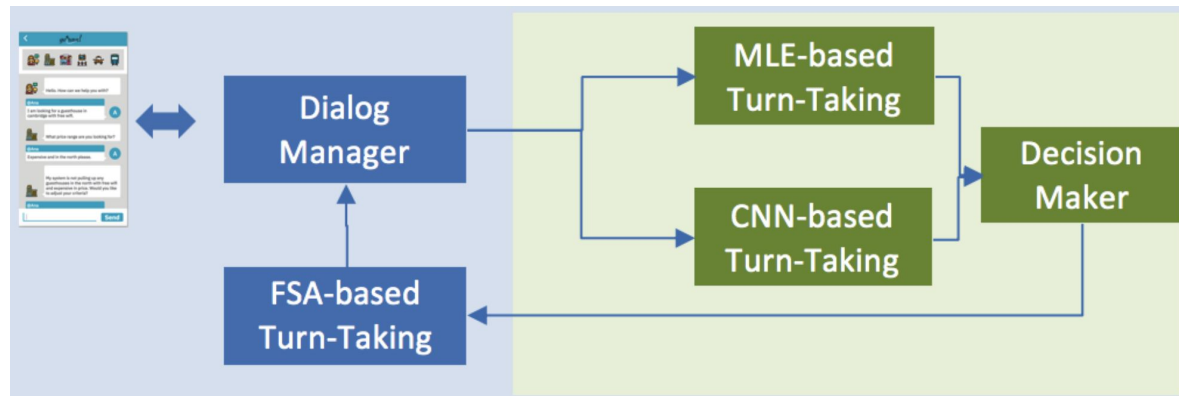
Section 3: Speaker Modeling

Representative Tasks

- **Turn-taking** is the process in which a participant in a conversation determines when it is appropriate to make an utterance and how that is accomplished
- Speaker segmentation, also known as speaker diarisation aims at finding speaker changing points in a conversation
- Speaker identification in novels is tasked to determine who says a quote in a given context by text analysis
- Utterance speaker search aims at searching for a speaker in the conversation history that shares the same speaker with the target utterance

Turn-taking

Hybrid of (a) **maximum likelihood estimation (MLE)** which encodes only the agent interaction order, and (b) **agent-and-content CNN** formats the **previous utterances** and the **agent names** as raw texts to predict the next speaker



$$s_{t+1} = \ell(x(t+1)) | \ell \in L$$
$$\ell(x(t+1)) = \begin{cases} \ell(x_1(t+1)) & \text{if } C_1 \geq k_1 \\ \ell(x_2(t+1)) & \text{if } C_2 < k_1 \\ & \text{and } C_2 \geq k_2 \\ travel_bot & \text{otherwise} \end{cases}$$

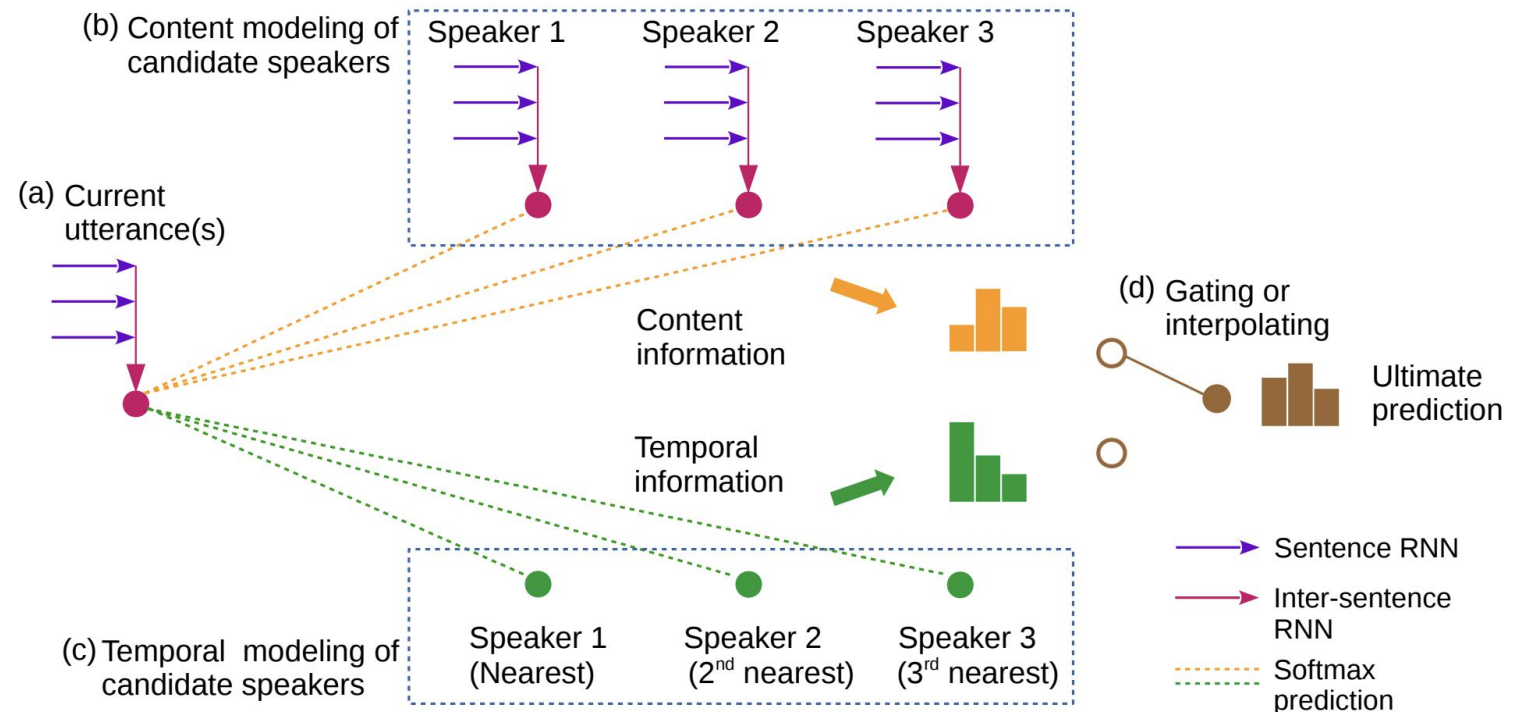
Representative Tasks

- Turn-taking is the process in which a participant in a conversation determines when it is appropriate to make an utterance and how that is accomplished
- **Speaker segmentation**, also known as speaker diarisation aims at finding speaker changing points in a conversation
- Speaker identification in novels is tasked to determine that who says a quote in a given context by text analysis
- Utterance speaker search aims at searching for a speaker in the conversation history that shares the same speaker with the target utterance

Speaker Segmentation

- Binary utterance-pair classification to judge **whether the speaker is changing before and after a point** based on the semantics discrepancy

- Select a speaker for **an MPC segment** given the speaker candidates



Zhao Meng, et al. *Hierarchical RNN with static sentence-level attention for text-based speaker change detection*. CIKM 2017.

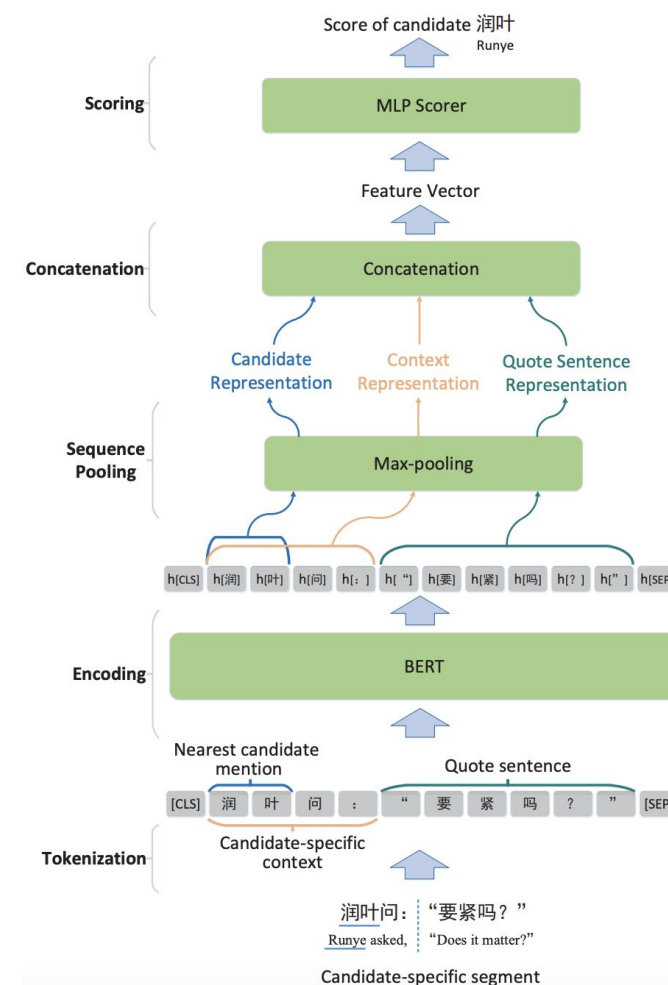
Zhao Meng, et al. *Towards neural speaker modeling in multi-party conversation: The task, dataset, and models*. LREC 2018.

Representative Tasks

- Turn-taking is the process in which a participant in a conversation determines when it is appropriate to make an utterance and how that is accomplished
- Speaker segmentation, also known as speaker diarisation aims at finding speaker changing points in a conversation
- **Speaker identification in novels** is tasked to determine that who says a quote in a given context by text analysis
- Utterance speaker search aims at searching for a speaker in the conversation history that shares the same speaker with the target utterance

Speaker Identification in Novels

- Formulate identification as a **scoring** task
- Candidate scoring network based on BERT **encode candidate-specific segments** to eliminate redundant context
- Post-revision based on the **speaker alternation pattern** in two-party dialogues

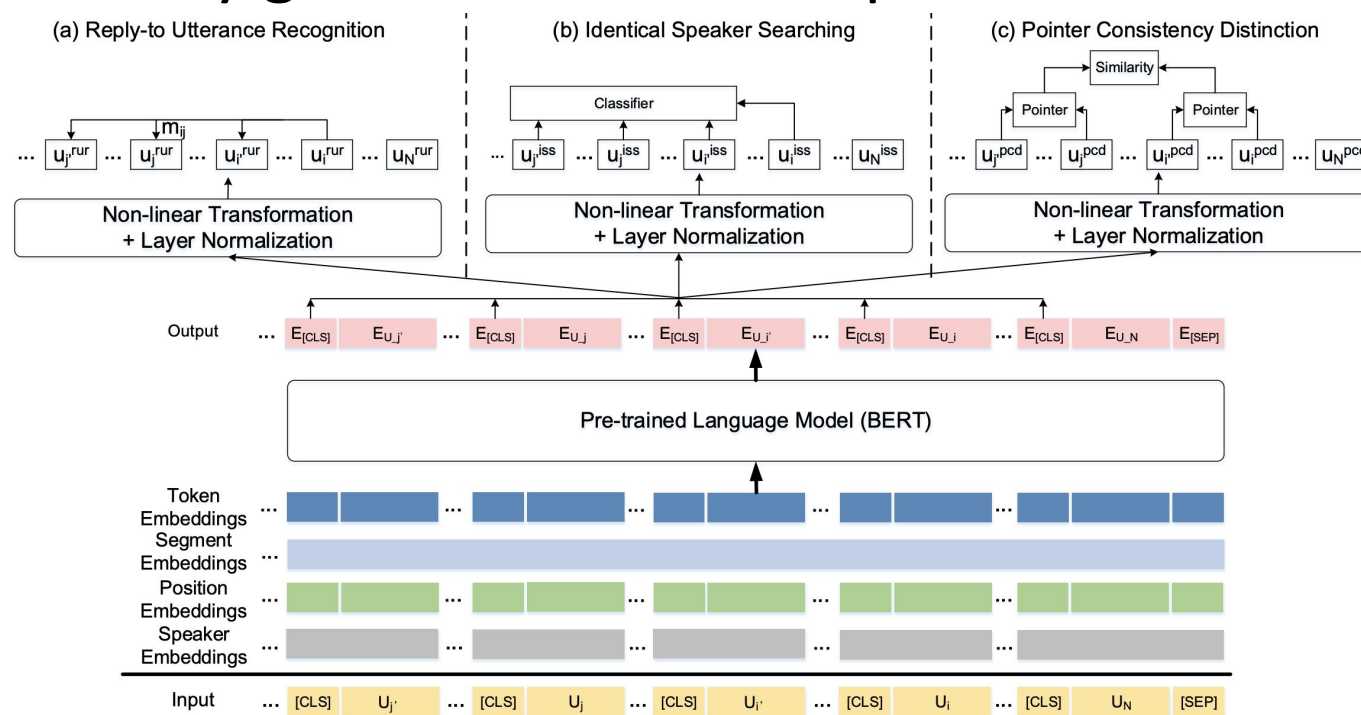


Representative Tasks

- Turn-taking is the process in which a participant in a conversation determines when it is appropriate to make an utterance and how that is accomplished
- Speaker segmentation, also known as speaker diarisation aims at finding speaker changing points in a conversation
- Speaker identification in novels is tasked to determine that who says a quote in a given context by text analysis
- **Utterance speaker search** aims at searching for a speaker in the conversation history that shares the same speaker with the target utterance

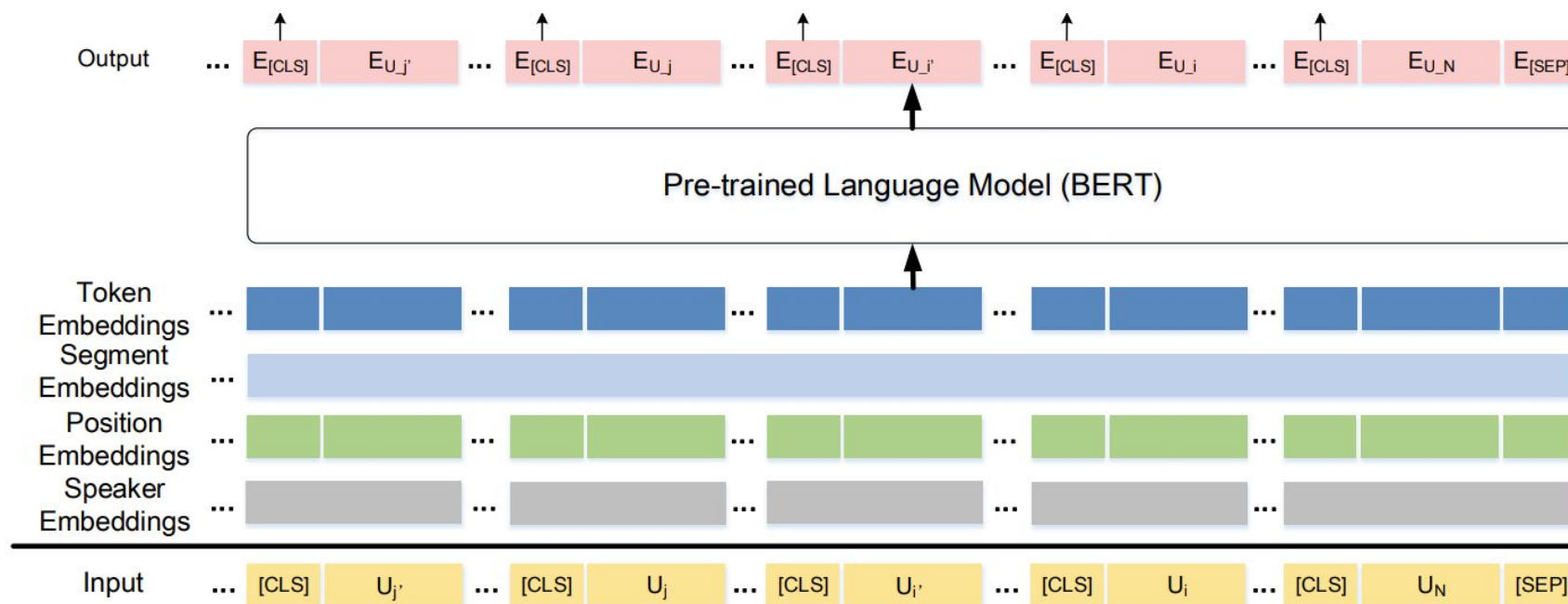
Utterance Speaker Search

Pretrain BERT with five **self-supervision** tasks, designed to model the underlying **interlocutor structure** and **utterance semantics**, which can be further effectively generalized to multiple MPC downstream tasks



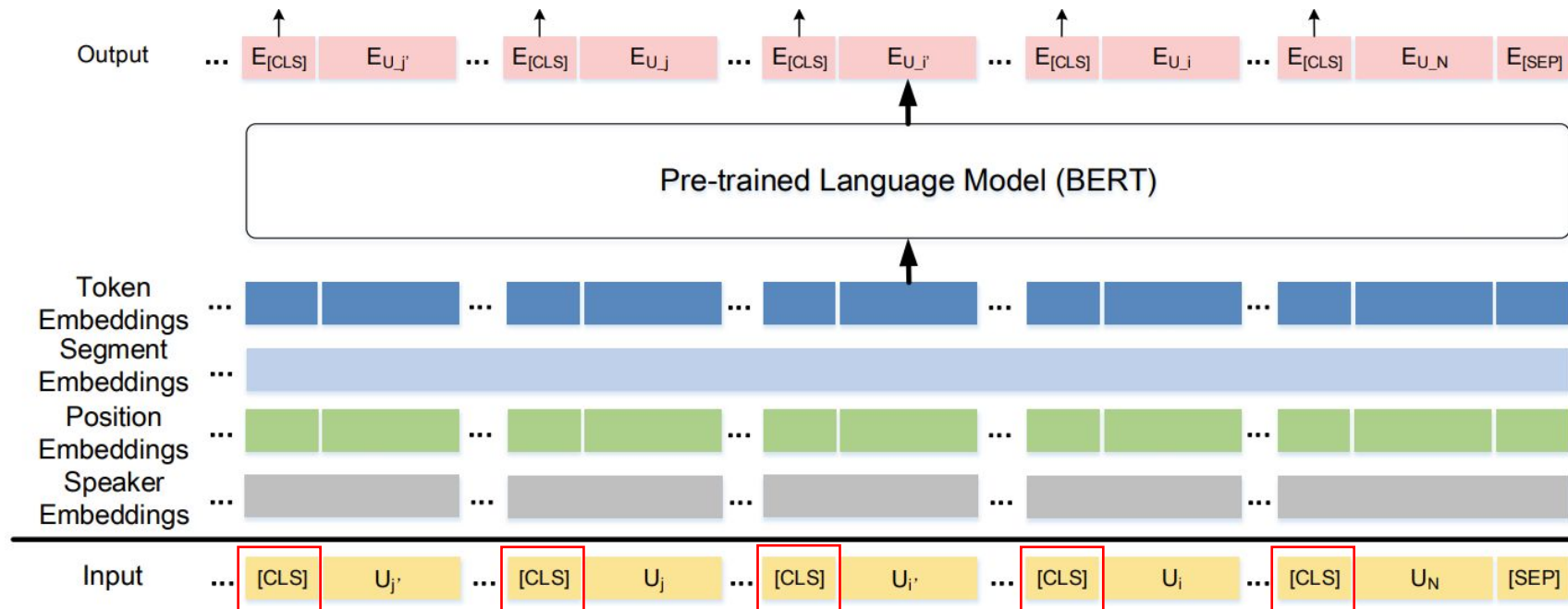
MPC-BERT: Model Overview

A [CLS] token is inserted at the start of each utterance



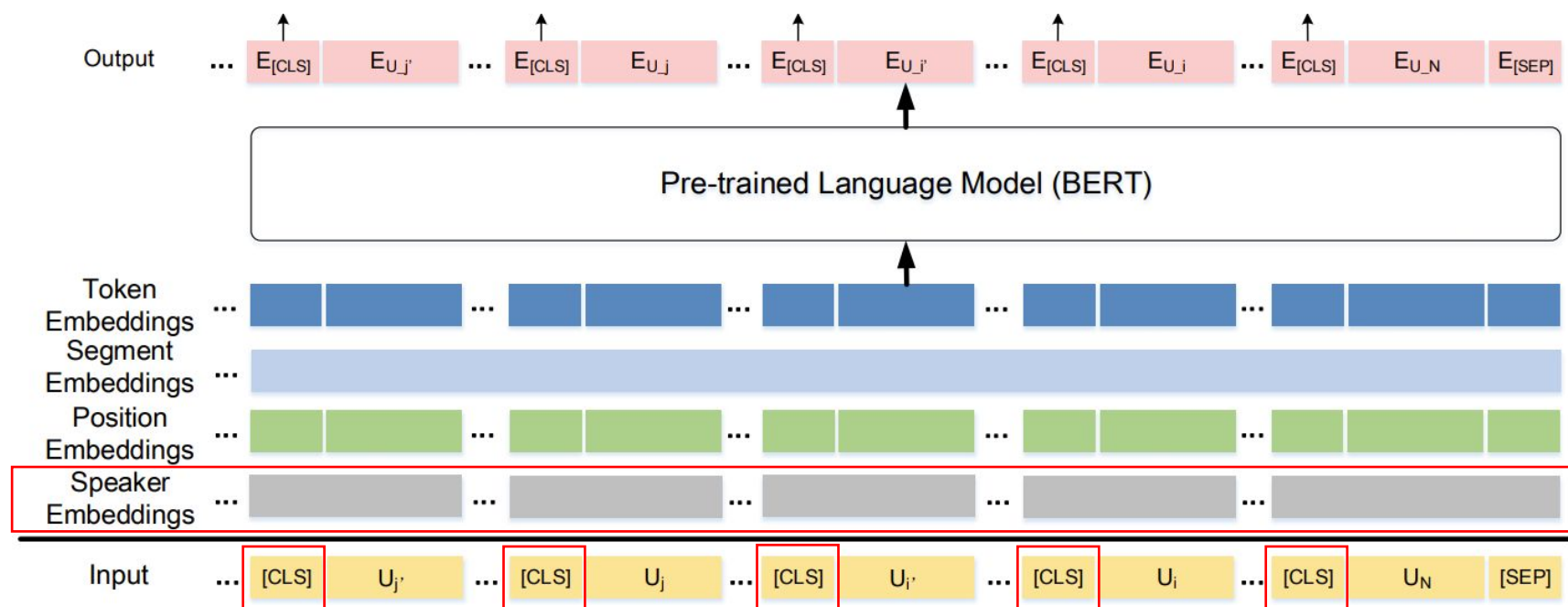
MPC-BERT: Model Overview

- A [CLS] token is inserted at the start of each utterance



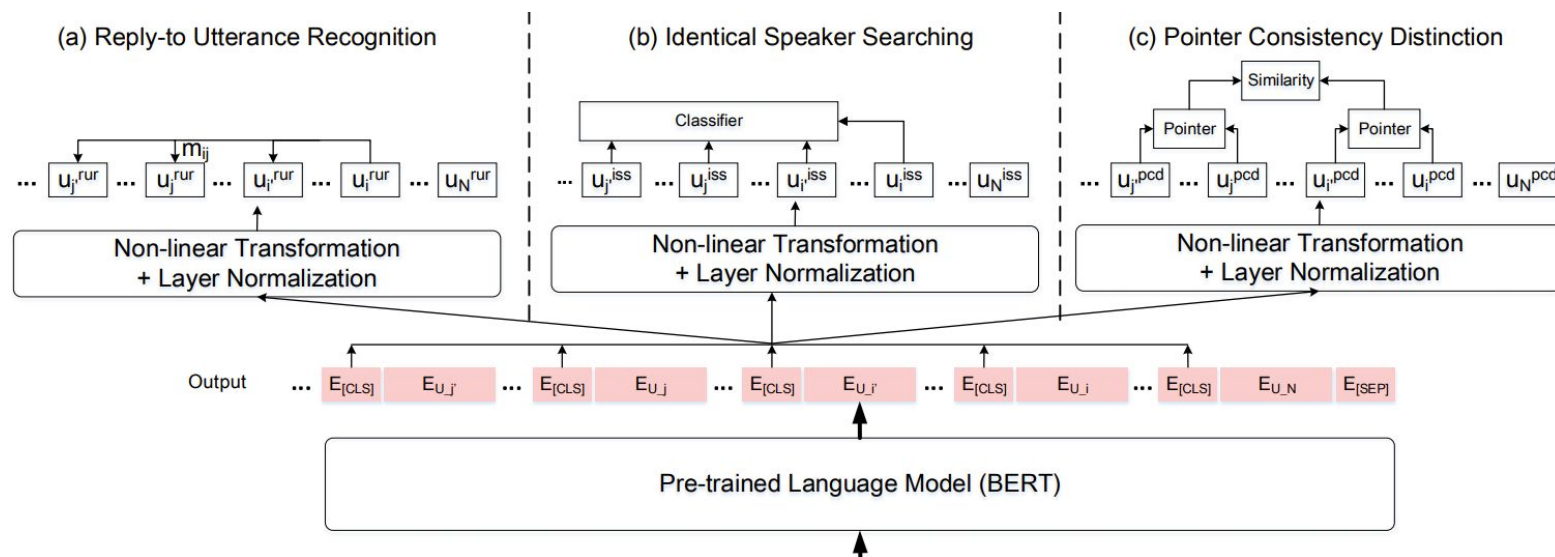
MPC-BERT: Model Overview

- A [CLS] token is inserted at the start of each utterance
- **Position-based speaker embeddings** (Gu et al., 2020) are introduced considering that interlocutors are inconsistent in different conversations



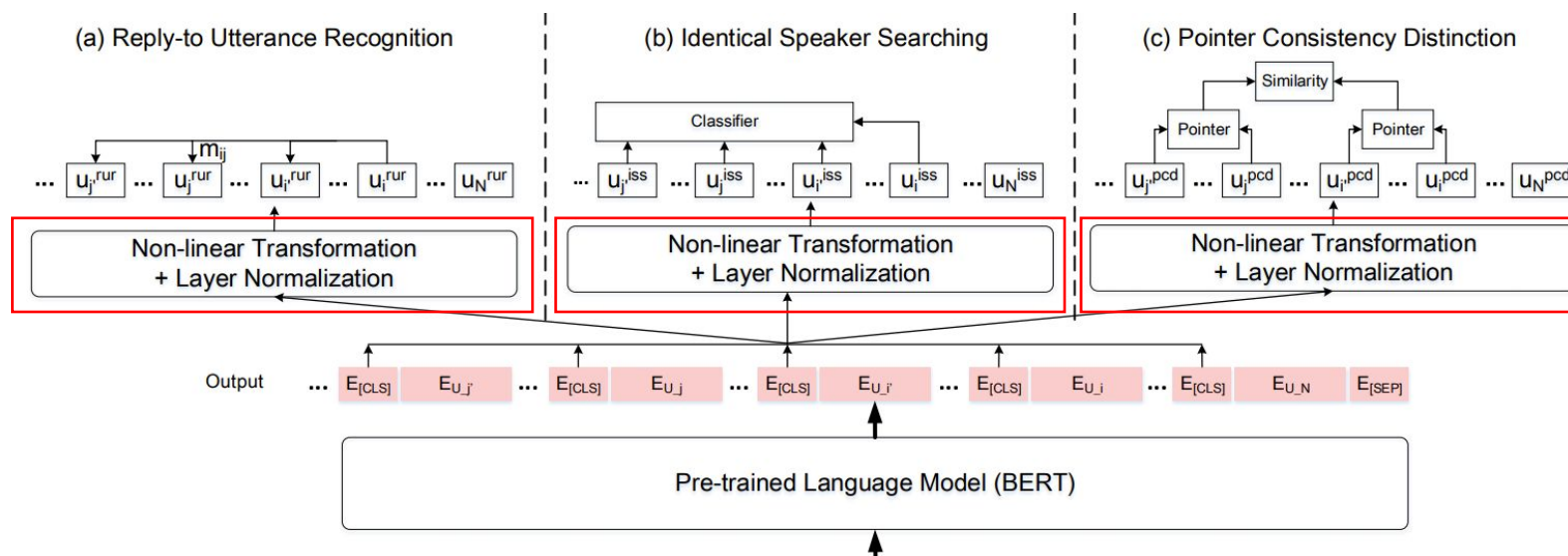
MPC-BERT: Interlocutor Structure Modeling

- Extract the **representations for each [CLS] token** representing utterances



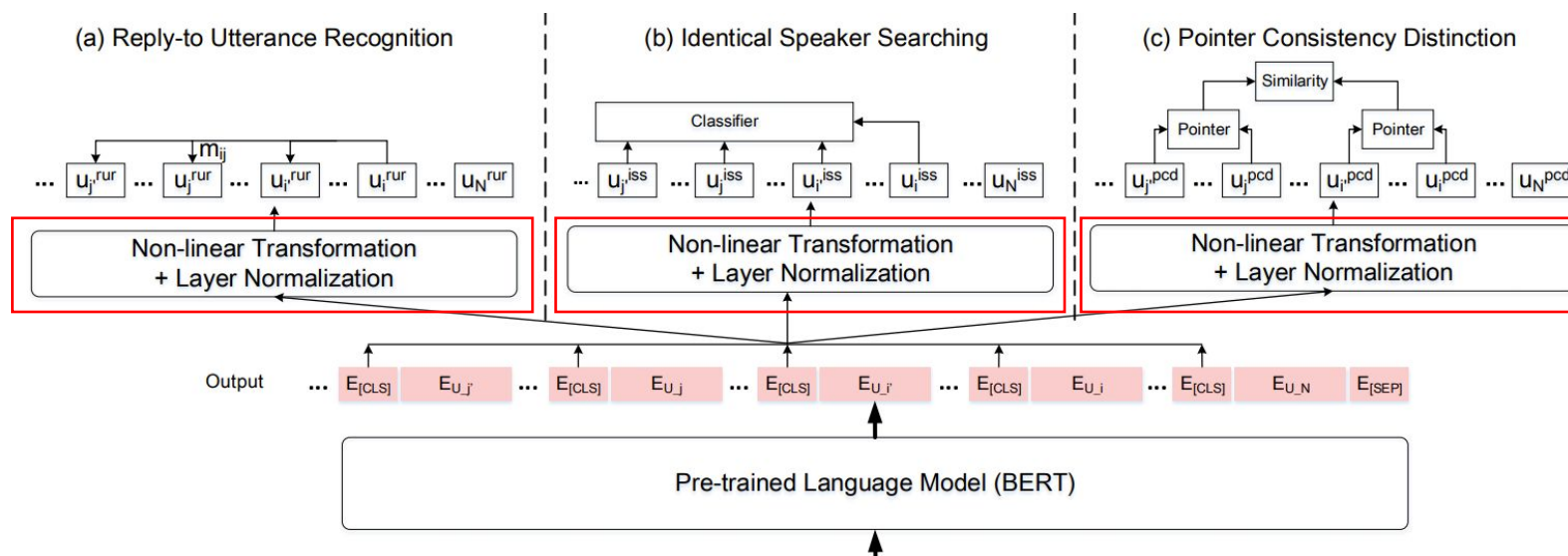
MPC-BERT: Interlocutor Structure Modeling

- Extract the **representations for each [CLS] token** representing utterances
- **Task-dependent non-linear transformations** are placed on top of BERT for three self-supervised tasks



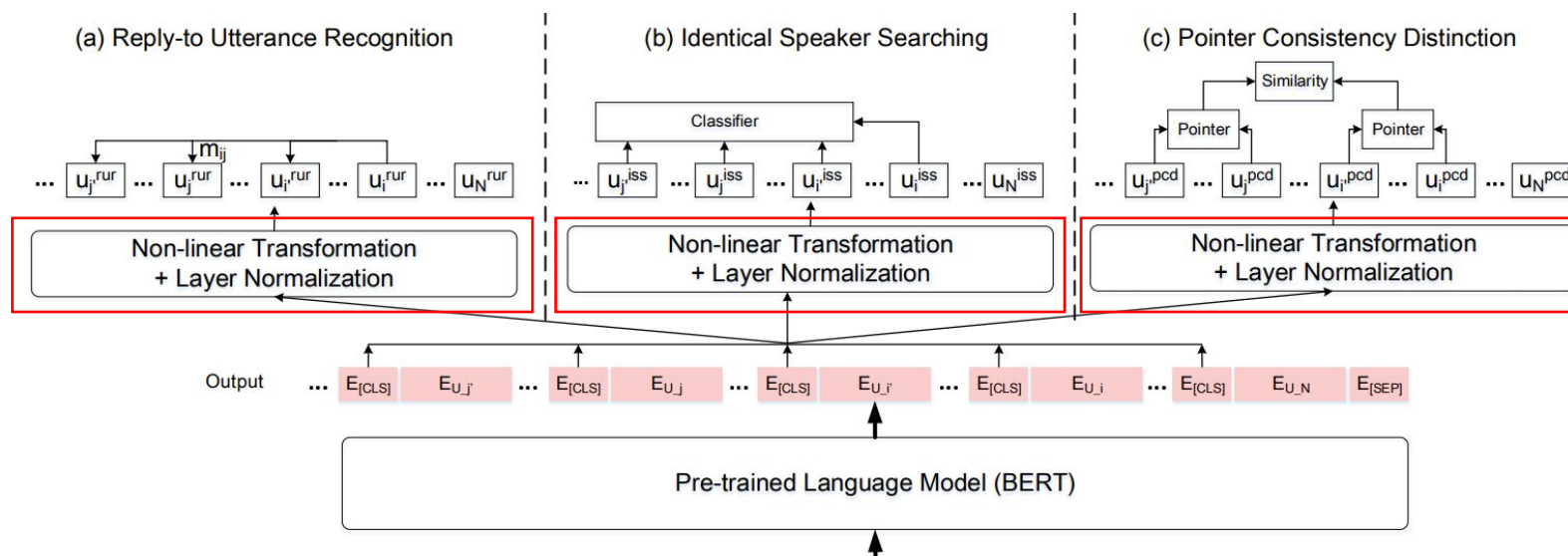
MPC-BERT: Interlocutor Structure Modeling

- Extract the **representations for each [CLS] token** representing utterances
- **Task-dependent non-linear transformations** are placed on top of BERT for three self-supervised tasks
- Encoding the input data only once is **computation-efficient**



MPC-BERT: Interlocutor Structure Modeling

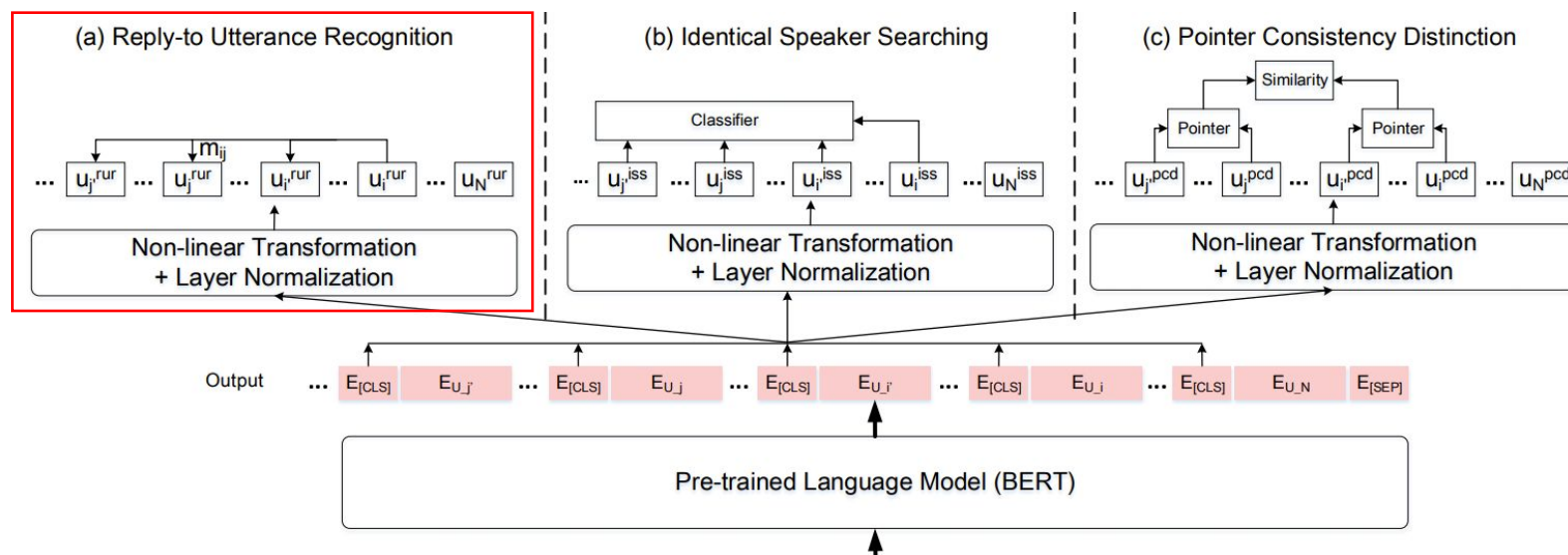
- Extract the **representations for each [CLS] token** representing utterances
- **Task-dependent non-linear transformations** are placed on top of BERT for three self-supervised tasks
- Encoding the input data only once is **computation-efficient**



Utterance semantics modeling part will be covered in Section

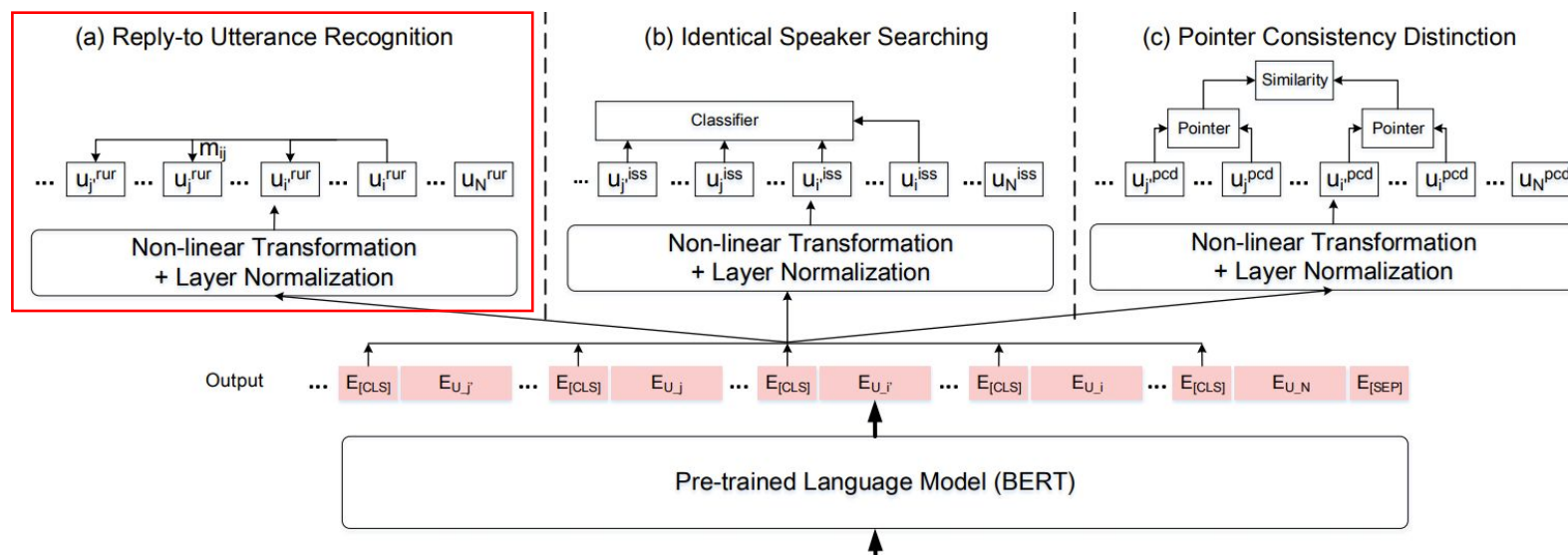
Reply-to Utterance Recognition

- **Motivation:** learn **which preceding utterance** the current utterance replies to



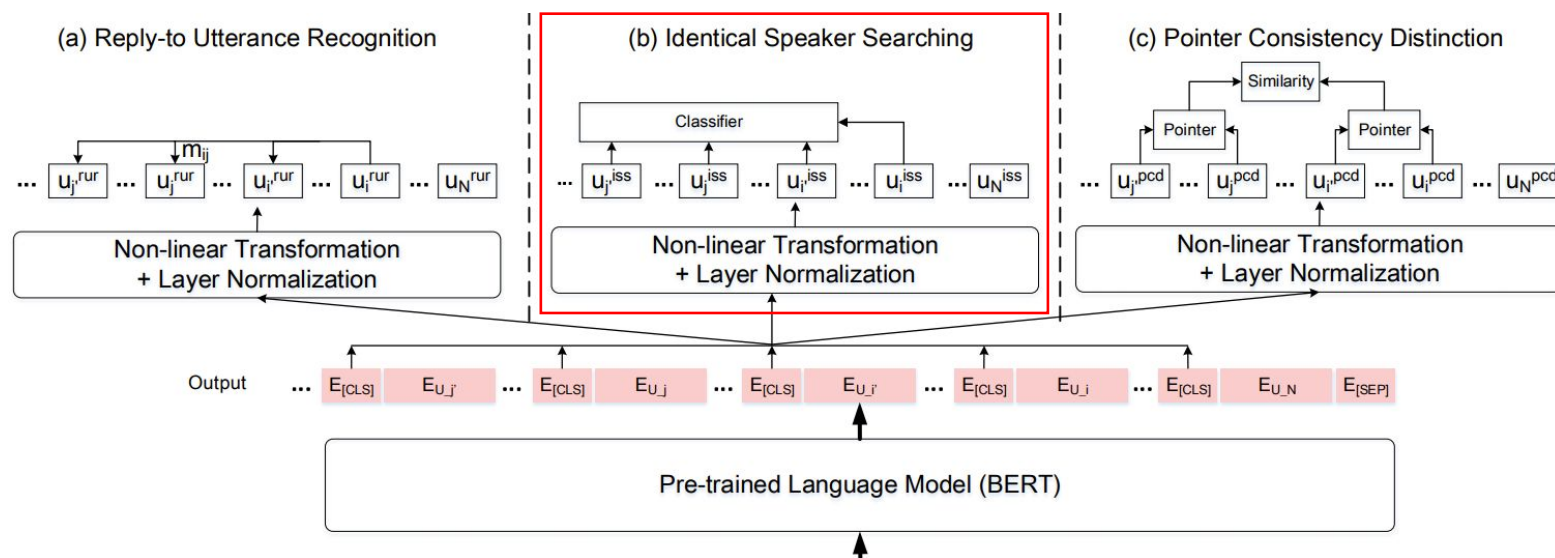
Reply-to Utterance Recognition

- **Motivation:** learn **which preceding utterance** the current utterance replies to
- **Implementation:** calculate the **matching scores** with all **its preceding utterances** for a target utterance



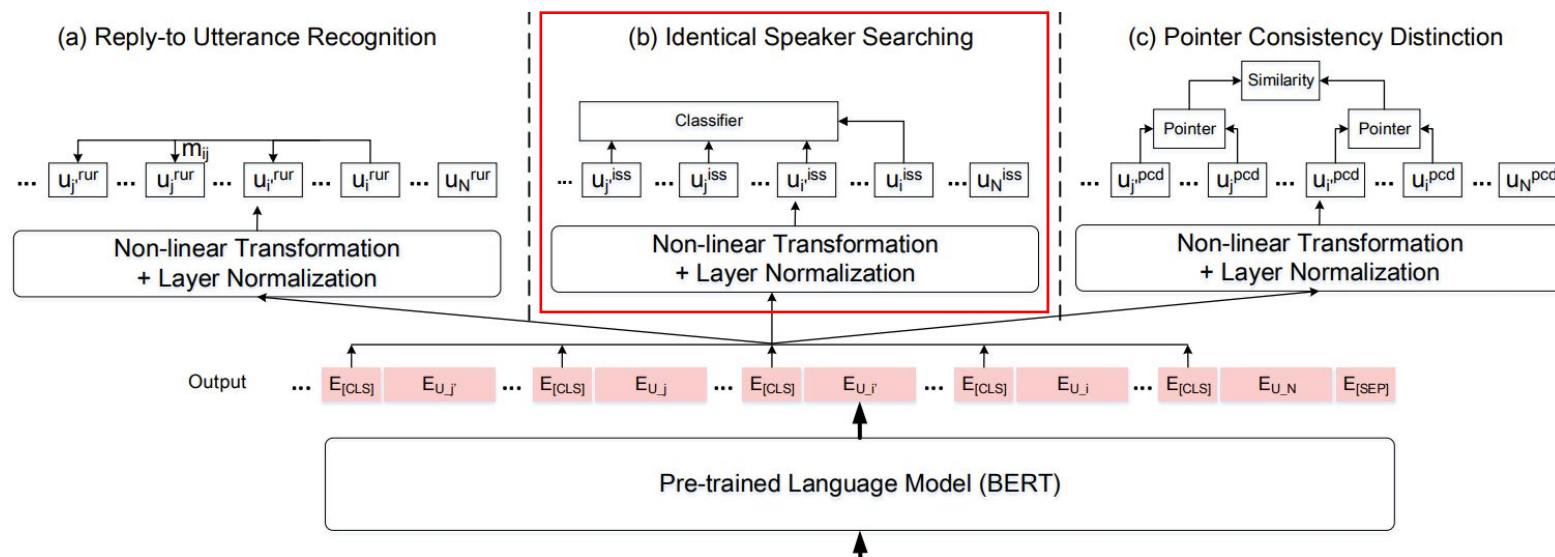
Identical Speaker Searching

- **Motivation:** reformulate as searching for the **utterances sharing the identical speaker**, since interlocutors **varies across conversations**



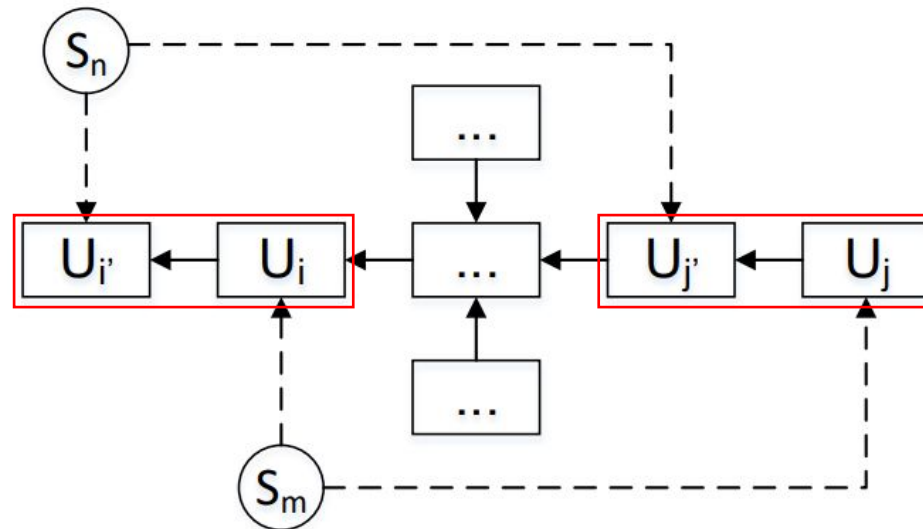
Identical Speaker Searching

- **Motivation:** reformulate as searching for the utterances sharing the identical speaker, since interlocutors varies across conversations
- **Implementation:** mask the speaker embedding of a target utterance, and calculate the probability of utterances sharing the same speaker



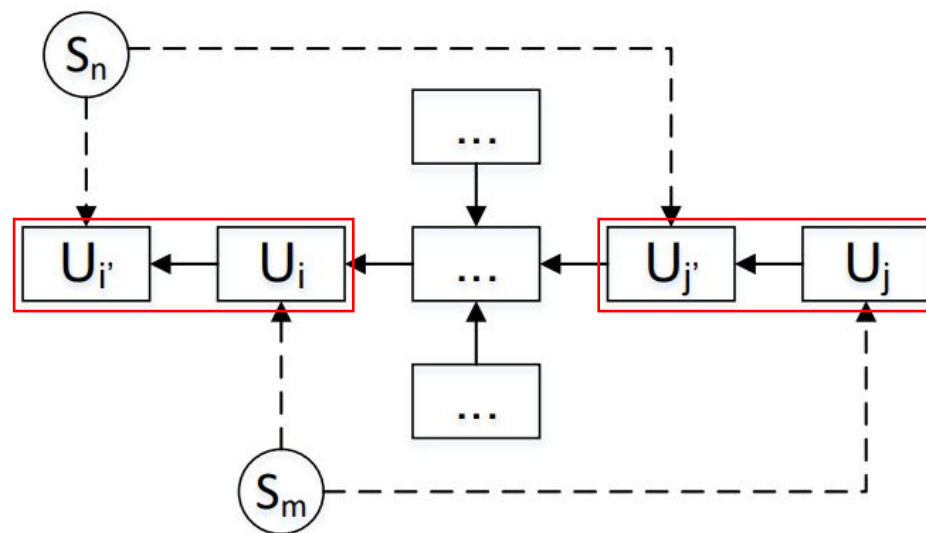
Pointer Consistency Distinction

- **Definition:** a **speaker-to-addressee pointer** is expressed as a **pair of utterances** representing the “**reply-to**” relationship



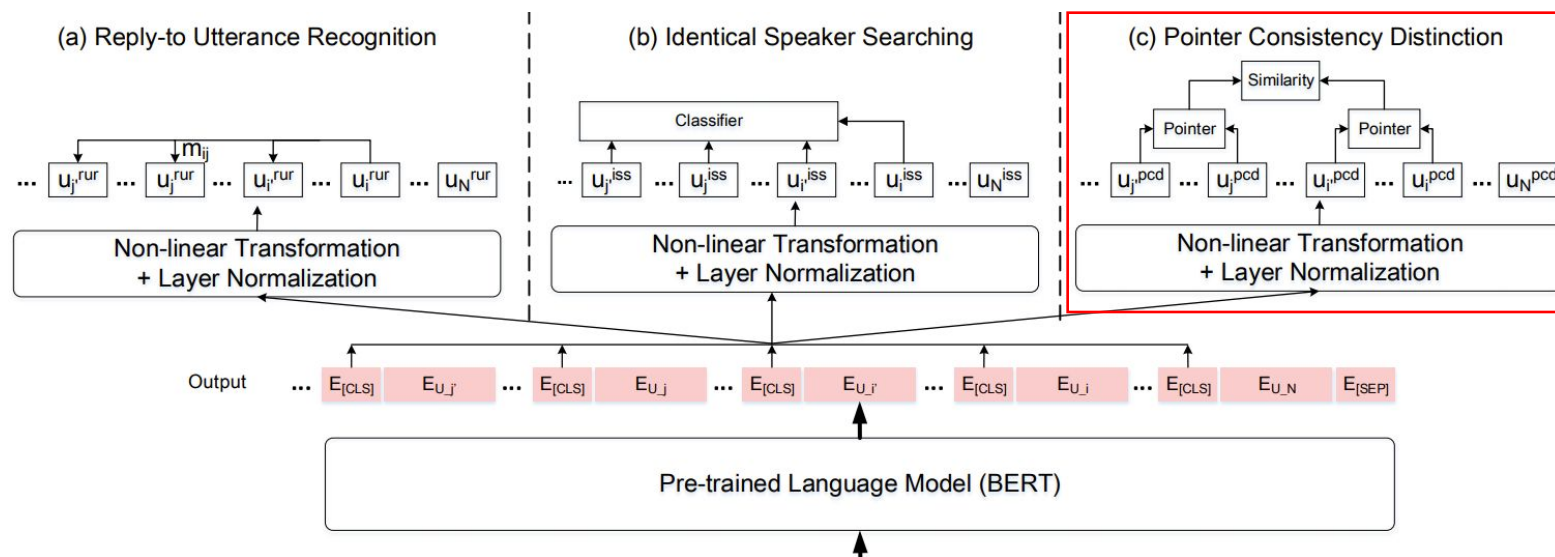
Pointer Consistency Distinction

- **Definition**: a **speaker-to-addressee pointer** is expressed as a pair of **utterances** representing the “**reply-to**” relationship
- **Assumption**: the representations of two pointers directing **from the same speaker to the same addressee** should be **consistent**



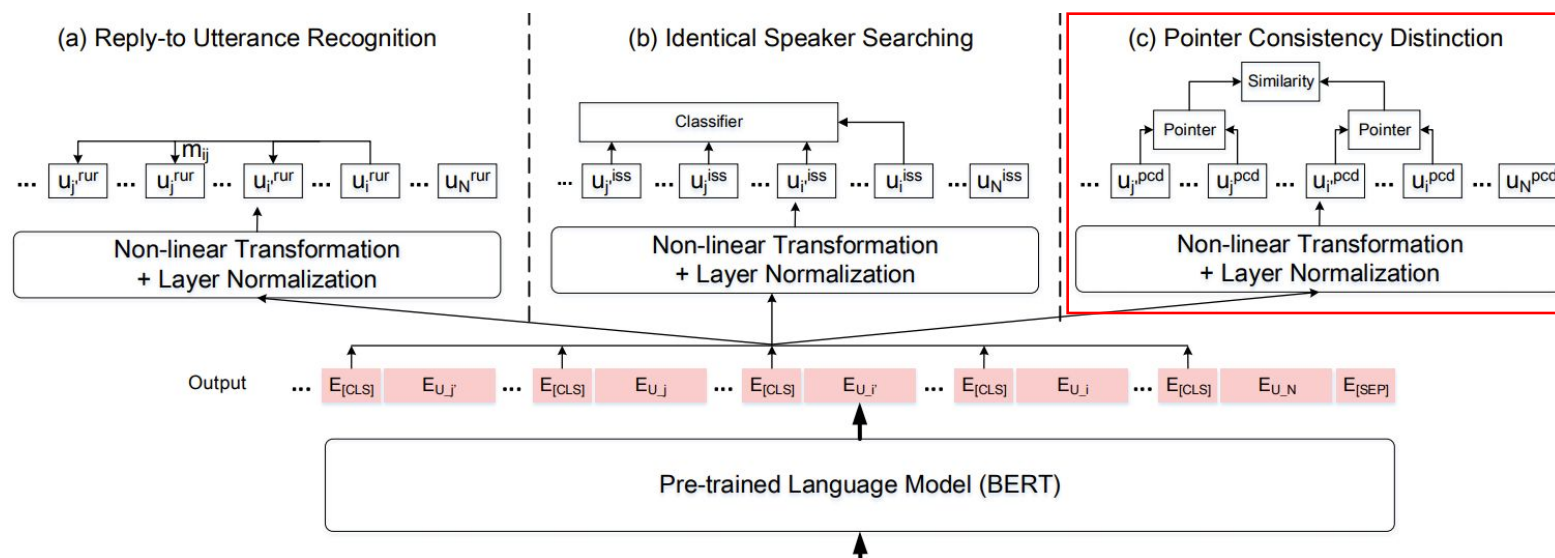
Pointer Consistency Distinction

- **Implementation** : a) capture the pointer information contained in each utterance pair



Pointer Consistency Distinction

- **Implementation** : a) capture the pointer information contained in each utterance pair
b) sample a consistent pointer and an inconsistent one from this conversation, and calculate similarities between every two pointers



Results

- **Metric:** Precision@1 (**P@1**)
- **Performance:** MPC-BERT outperforms SA-BERT by margins of **7.66%**, **2.60%**, **3.38%** and **4.24%** respectively in terms of P@1
- **Ablation:** **ISS** and **RUR** contribute the most

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
BERT (Devlin et al., 2019)	71.81	62.24	53.17	51.58
SA-BERT (Gu et al., 2020a)	75.88	64.96	57.62	54.28
MPC-BERT	83.54	67.56	61.00	58.52
MPC-BERT w/o. RUR	82.48	66.88	60.12	57.33
MPC-BERT w/o. ISS	77.95	66.77	60.03	56.73
MPC-BERT w/o. PCD	83.39	67.12	60.62	58.00
MPC-BERT w/o. MSUR	83.51	67.21	60.76	58.03
MPC-BERT w/o. SND	83.47	67.04	60.44	58.12

Table 4: Evaluation results of speaker identification on the test sets in terms of P@1. Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p -value < 0.05).

Section 4: Addressee Modeling

Representative Tasks

- **Addressee recognition** is tasked to **directly recognize** the addressee of target utterances given the interlocutor set in this conversation (**explicit** addressee modeling)
- Dialogue disentanglement aims at disentangling a whole conversation from a data stream into several threads via the underlying reply relationships, so that each thread is about a specific topic (implicit addressee modeling)

Addressee Recognition

Speaker	Utterance	Addressee
User 1	"Good point, tmux is the thing I miss."	–
User 1	"Cool thanks for ur help." @User 4	User 4
User 2	"Ahha, you r using something like cpanel."	–
User 3	"Yeah 1.4.0 exactly." @User 2	User 2
User 4	"my pleasure :)"	–

Addressee Recognition

Speaker	Utterance	Addressee
User 1	"Good point, tmux is the thing I miss."	–
User 1	"Cool thanks for ur help." @User 4	User 4
User 2	"Ahha, you r using something like cpanel."	–
User 3	"Yeah 1.4.0 exactly." @User 2	User 2
User 4	"my pleasure :)"	–

Not all the addressees are specified!

Addressee Recognition

Speaker	Utterance	Addressee
User 1	"Good point, tmux is the thing I miss."	–
User 1	"Cool thanks for ur help." @User 4	User 4
User 2	"Ahha, you r using something like cpanel."	–
User 3	"Yeah 1.4.0 exactly." @User 2	User 2
User 4	"my pleasure :)"	–

Not all the addressees are specified!

- **Target at only the last utterance:** Ouchi and Tsuboi (2016); Zhang et al. (2018); Gu et al. (2023); Zhu et al. (2023)

Addressee Recognition

Speaker	Utterance	Addressee
User 1	"Good point, tmux is the thing I miss."	–
User 1	"Cool thanks for ur help." @User 4	User 4
User 2	"Ahha, you r using something like cpanel."	–
User 3	"Yeah 1.4.0 exactly." @User 2	User 2
User 4	"my pleasure :)"	–

Not all the addressees are specified!

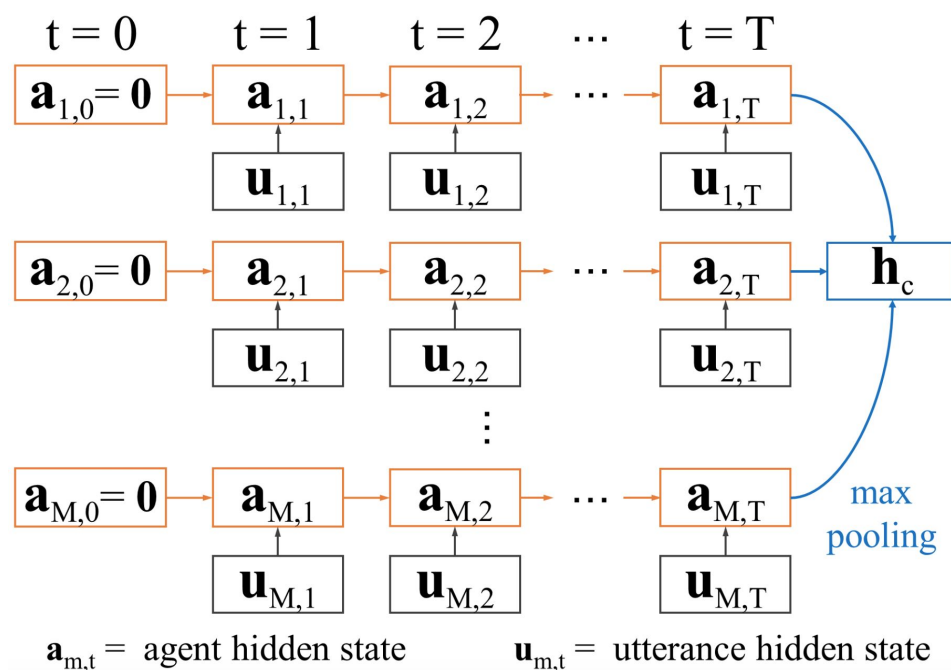
- **Target at only the last utterance:** Ouchi and Tsuboi (2016); Zhang et al. (2018); Gu et al. (2023); Zhu et al. (2023)
- **Target at all utterances where the addressees are missing:** Le et al. (2019); Gu et al. (2021)

Dynamic RNN

The **agent representation changes** along with each time step

Dynamic RNN

The **agent representation changes** along with each time step

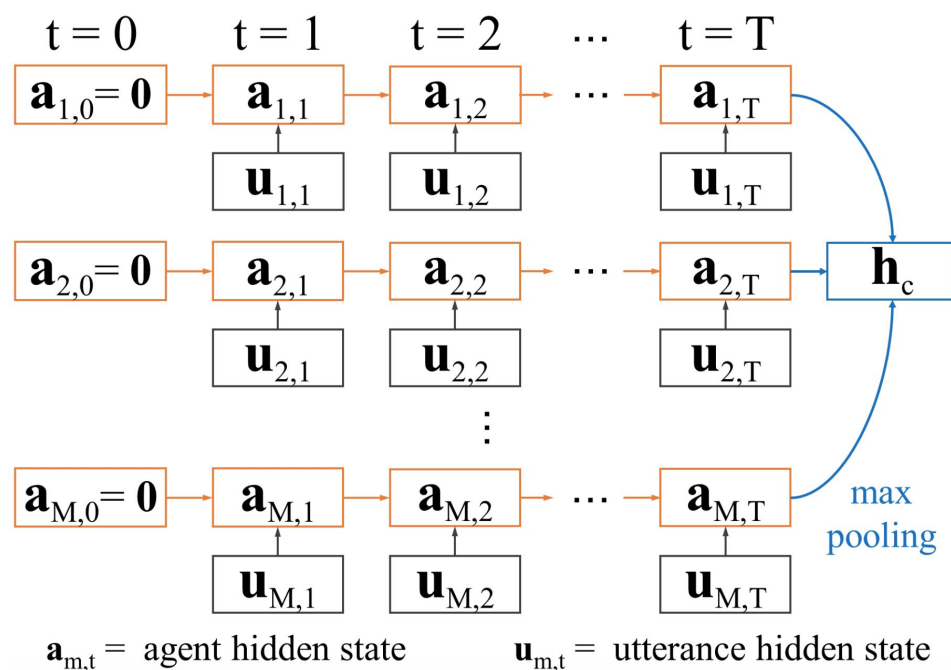


The states of the agents that

- are **speaking** at the time are updated by **consuming the utterance vector**

Dynamic RNN

The **agent representation changes** along with each time step



The states of the agents that

- are **speaking** at the time are updated by **consuming the utterance vector**
- are **not speaking** at the time are updated by **zero vectors**

Speaker Interaction RNN

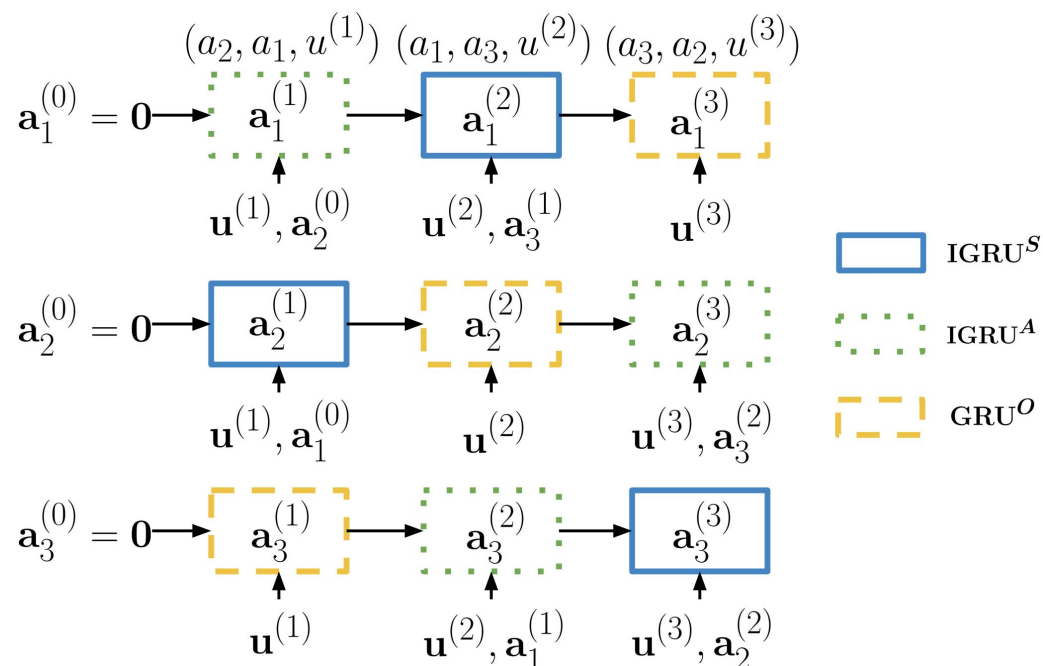
Interlocutors play **different roles** (sender, addressee, observer) which vary across turns

Speaker Interaction RNN

Interlocutors play **different roles** (sender, addressee, observer) which vary across turns □ update interlocutor embeddings **role-sensitively**

Speaker Interaction RNN

Interlocutors play **different roles** (sender, addressee, observer) which vary across turns □ update interlocutor embeddings **role-sensitively**



The same interlocutor embedding table is updated in **different units depending on the role**

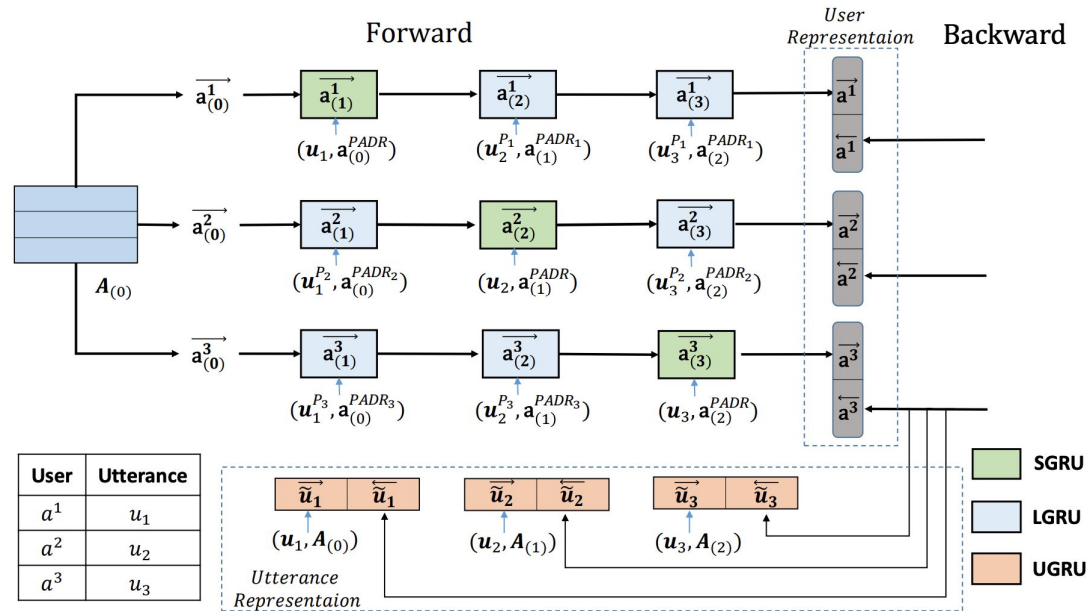
- IGRU^S for sender
- IGRU^A for addressee
- GRU^O for observer

Who-to-Whom

Identify **all the missing addressees** in a conversation session and model interlocutors and utterances **jointly** and **interactively**

Who-to-Whom

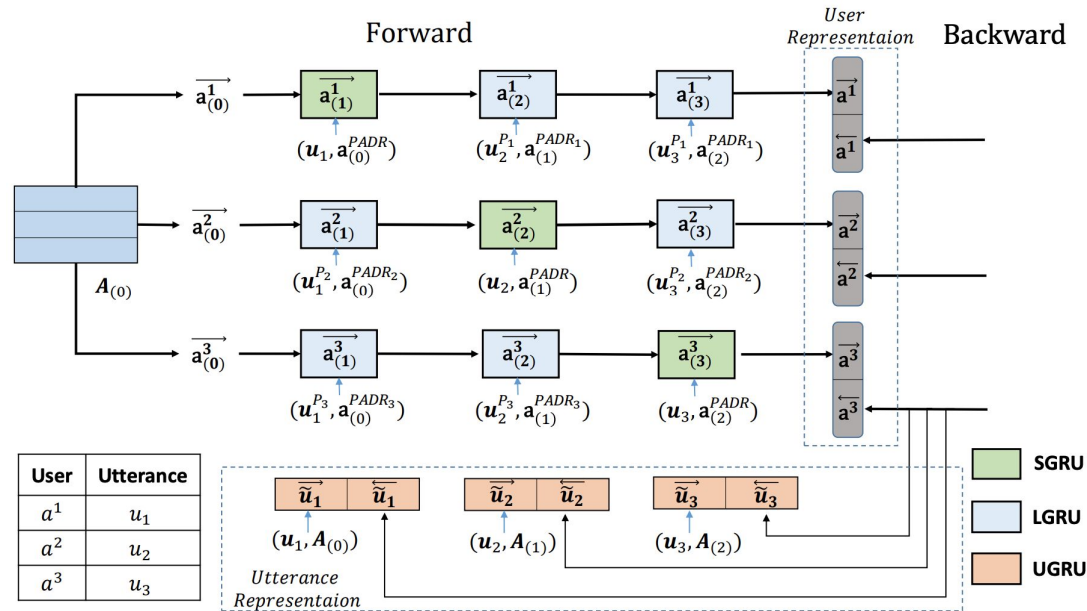
Identify **all the missing addressees** in a conversation session and model interlocutors and utterances **jointly** and **interactively**



- Track users' states with utterance embeddings (DRNN, SI-RNN), i.e., **utterance -> user**

Who-to-Whom

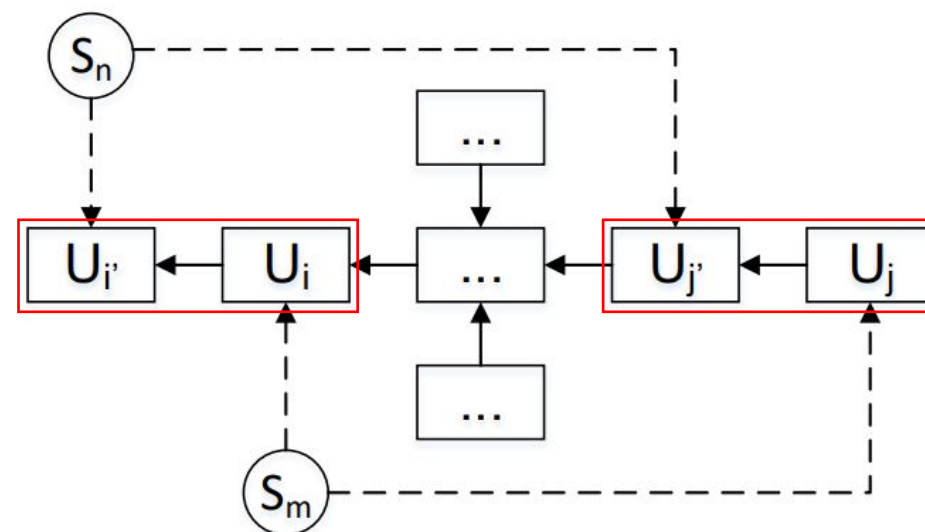
Identify **all the missing addressees** in a conversation session and model interlocutors and utterances **jointly** and **interactively**



- Track users' states with utterance embeddings (DRNN, SI-RNN), i.e., **utterance -> user**
- Fuse users' states into the utterance embeddings, i.e., **user -> utterance**

MPC-BERT

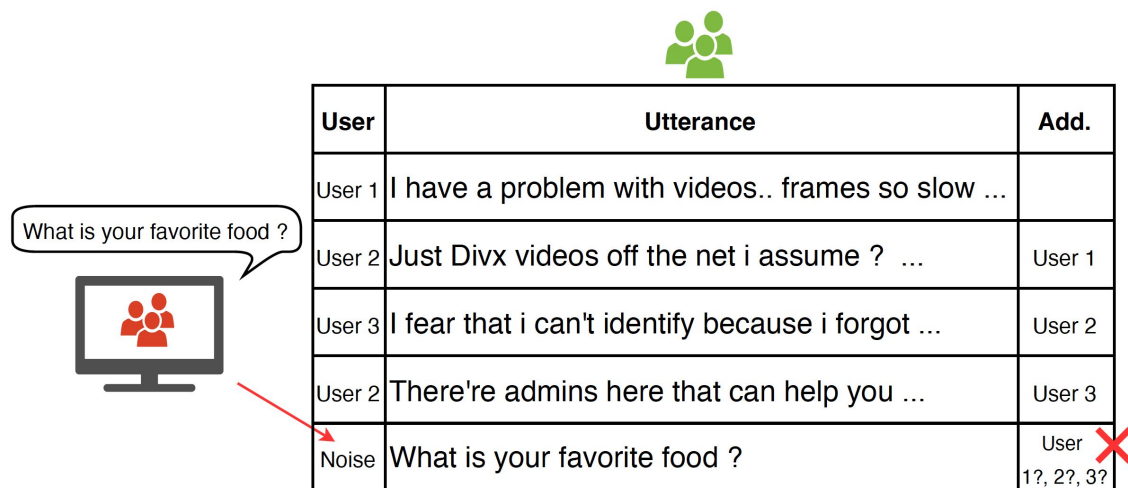
- **Identical speaker searching:** to searching for the utterances sharing the identical speaker
- **Reply-to utterance recognition:** to learn which preceding utterance the current utterance replies to
- **Pointer consistency distinction:** a speaker-to-addressee pointer is defined. Assume that representations of pointers directing from the same speaker to the same addressee should be consistent



Discussed in Section 3!

RARM

Focus on **robust** addressee recognition, where the noise perturbations are **semantically complete**, but are **not intended** for the conversation



The diagram illustrates a multi-party conversation scenario. A green icon of three people is positioned above a table. To the left of the table, a computer monitor icon displays a red icon of three people, with a speech bubble above it containing the text "What is your favorite food ?". A red arrow points from this speech bubble to the "Noise" row of the table. The table has three columns: "User", "Utterance", and "Add.". The rows are as follows:

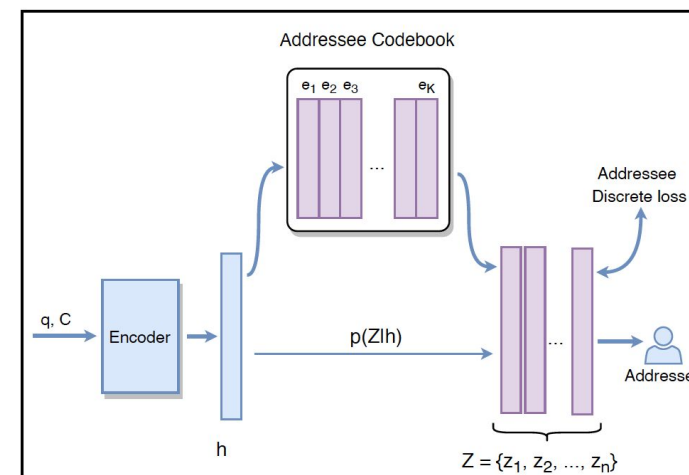
User	Utterance	Add.
User 1	I have a problem with videos.. frames so slow ...	
User 2	Just Divx videos off the net i assume ? ...	User 1
User 3	I fear that i can't identify because i forgot ...	User 2
User 2	There're admins here that can help you ...	User 3
Noise	What is your favorite food ?	User 1?, 2?, 3? 

RARM

Focus on **robust** addressee recognition, where the noise perturbations are **semantically complete**, but are **not intended** for the conversation



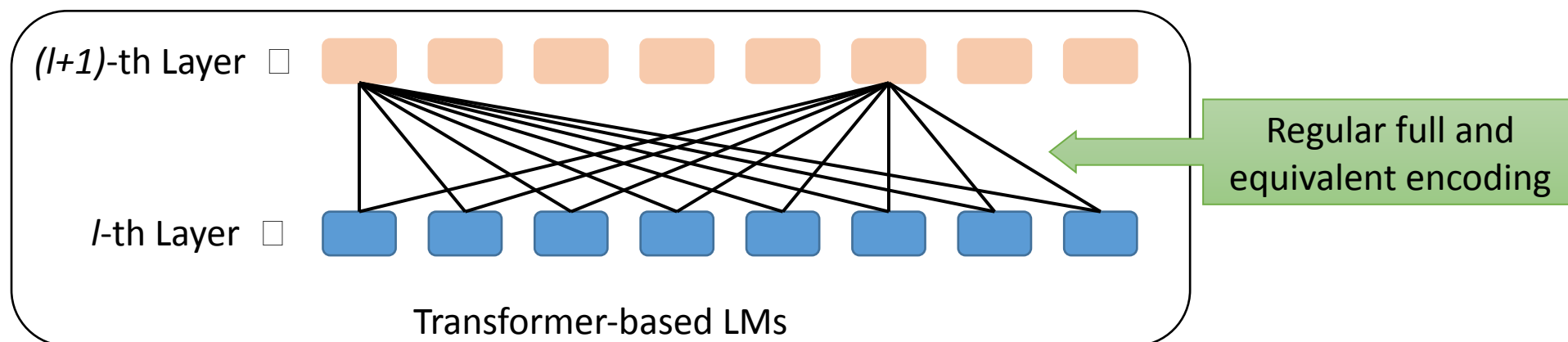

User	Utterance	Add.
User 1	I have a problem with videos.. frames so slow ...	
User 2	Just Divx videos off the net i assume ? ...	User 1
User 3	I fear that i can't identify because i forgot ...	User 2
User 2	There're admins here that can help you ...	User 3
Noise	What is your favorite food ?	User 1?, 2?, 3? 



Discretize addressees into a codebook with VQ-VAE to solve the issue of **unknown number of addressees** in noisy environment

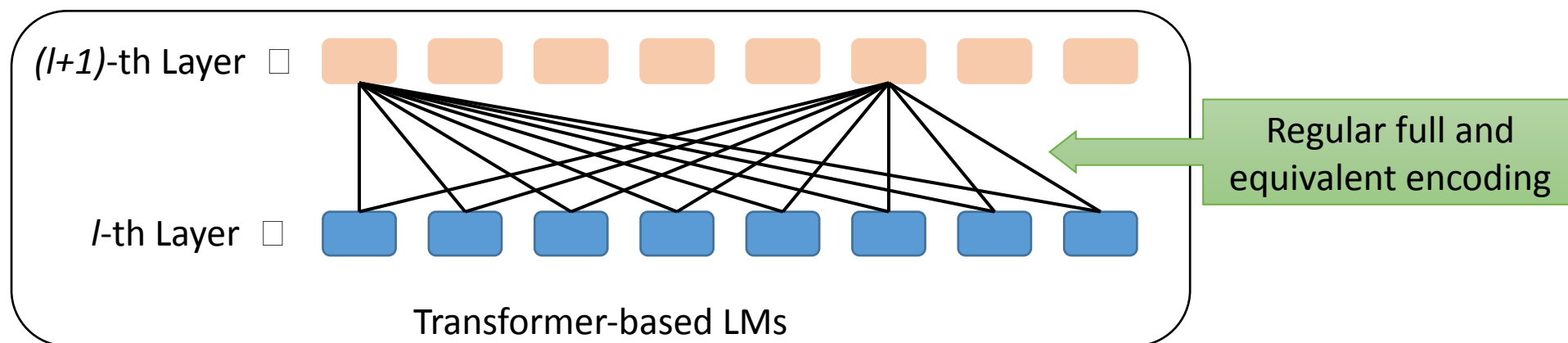
GIFT

- **Motivation:** the **full and equivalent connections** among utterance tokens ignore the **sparse but distinctive dependency** of one utterance on another



GIFT

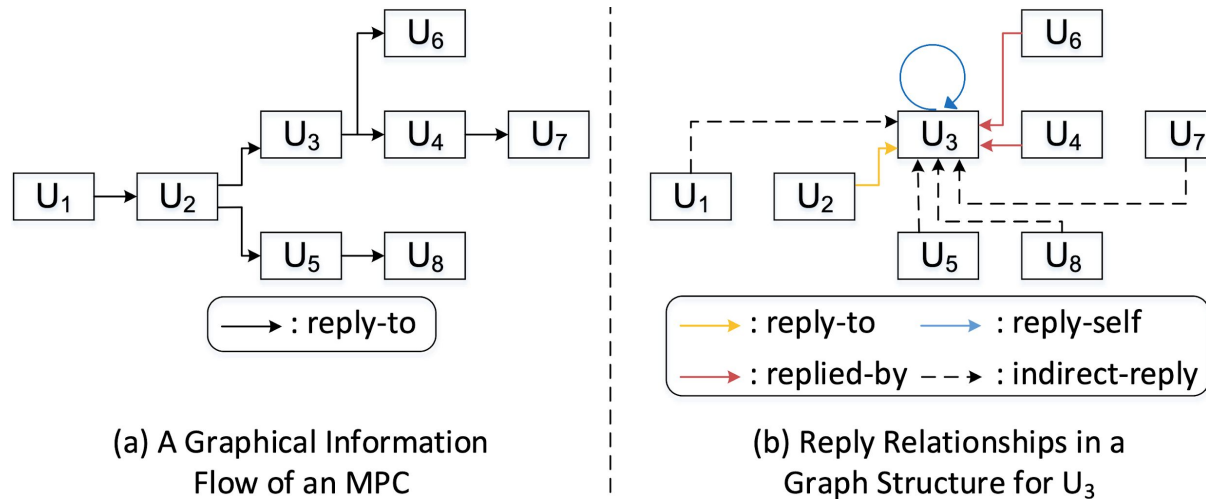
- **Motivation:** the **full and equivalent connections** among utterance tokens ignore the **sparse but distinctive dependency** of one utterance on another



- **Methodology:** to distinguish different **utterance relationships** for modeling the inherent **MPC graph structure** via graph-induced fine-tuning

GIFT Graph Topology

Four types of edges: *reply-to*, *replied-by*, *reply-self* and *indirect-reply* are designed to distinguish different relationships between utterances



* Rectangles ($\boxed{U}$) denote utterances, and solid lines ($\longrightarrow$) represent the "reply" relationship between two utterances

Graph-Induced Signals Integration

- Integrated in the **attention mechanism** by utilizing **edge-type-dependent parameters** to **refine** the attention weights

$$\text{Atten}(q, k, v) = \text{softmax}(\phi(e_{q,v}) \frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}}) \mathbf{v}$$

where $e_{q,v} \in \{\text{reply-to}, \text{replied-by}, \text{reply-self}, \text{indirect-reply}\}$

Graph-Induced Signals Integration

- Integrated in the **attention mechanism** by utilizing **edge-type-dependent parameters** to **refine** the attention weights

$$\text{Atten}(q, k, v) = \text{softmax}(\phi(e_{q,v}) \frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}}) \mathbf{v}$$

where $e_{q,v} \in \{\text{reply-to}, \text{replied-by}, \text{reply-self}, \text{indirect-reply}\}$

- **reply-to**: what the current utterance should be like given the **prior utterance it replies to**

Graph-Induced Signals Integration

- Integrated in the **attention mechanism** by utilizing **edge-type-dependent parameters** to **refine** the attention weights

$$\text{Atten}(q, k, v) = \text{softmax}(\phi(e_{q,v}) \frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}}) \mathbf{v}$$

where $e_{q,v} \in \{\text{reply-to}, \text{replied-by}, \text{reply-self}, \text{indirect-reply}\}$

- **reply-to**: what the current utterance should be like given the **prior utterance it replies to**
- **replied-by**: how the **posterior utterances** amend the modeling of the current utterance

Graph-Induced Signals Integration

- Integrated in the **attention mechanism** by utilizing **edge-type-dependent parameters** to **refine** the attention weights

$$\text{Atten}(q, k, v) = \text{softmax}(\phi(e_{q,v}) \frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}}) \mathbf{v}$$

where $e_{q,v} \in \{\text{reply-to}, \text{replied-by}, \text{reply-self}, \text{indirect-reply}\}$

- **reply-to**: what the current utterance should be like given the **prior utterance it replies to**
- **replied-by**: how the **posterior utterances** amend the modeling of the current utterance
- **reply-self**: how much of the **original semantics** should be kept

Graph-Induced Signals Integration

- Integrated in the **attention mechanism** by utilizing **edge-type-dependent parameters** to **refine** the attention weights

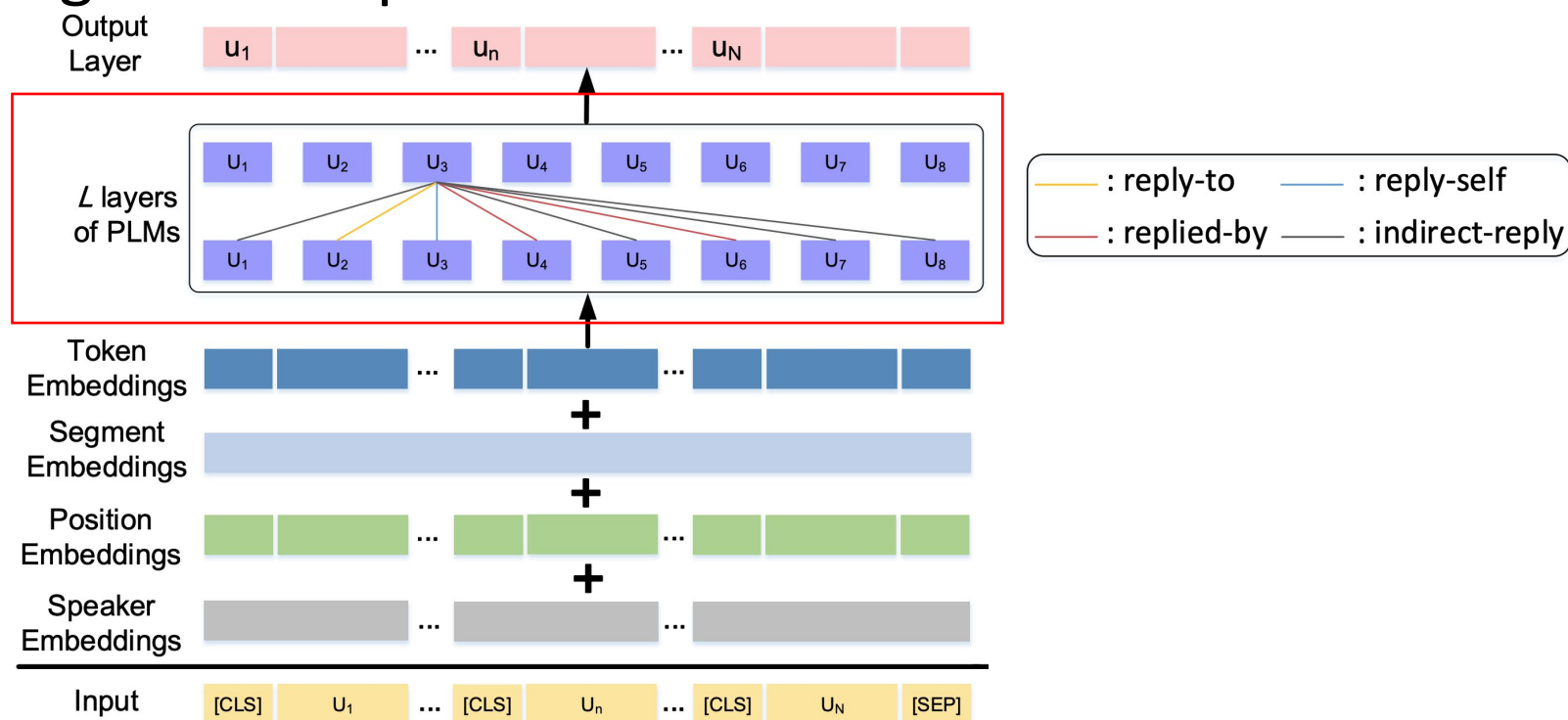
$$\text{Atten}(q, k, v) = \text{softmax}(\phi(e_{q,v}) \frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}}) \mathbf{v}$$

where $e_{q,v} \in \{\text{reply-to}, \text{replied-by}, \text{reply-self}, \text{indirect-reply}\}$

- **reply-to**: what the current utterance should be like given the **prior utterance it replies to**
- **replied-by**: how the **posterior utterances** amend the modeling of the current utterance
- **reply-self**: how much of the **original semantics** should be kept
- **indirect-reply**: connect **the rest of the utterances** for contextualization

GIFT Overview

Input data following MPC-BERT that (1) inserts **[CLS] tokens** at the start of each utterance, and (2) introduces **position-based speaker embeddings** to distinguish the speakers of utterances



Why These Edges Work?

- Consider both **semantic similarity** and **structural relationships** between two utterance tokens
- Distinguish **different relationships** between utterances, and model **utterance dependency** following the **graph-induced topology** for better contextualized encoding
- Characterize **fine-grained interactions** during LM internal encoding
- Reflect **graphical conversation structure and flow** in Transformer

Results: Addressee Recognition

GIFT improves **BERT** by margins of **2.92%, 2.73%, 5.75% and 5.08%** on these test sets respectively in terms of **Precision (P@1)**

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
Preceding (Le et al., 2019)	-	55.73	55.63	55.62
SRNN (Ouchi and Tsuboi, 2016)	-	60.26	60.66	60.98
SHRNN (Serban et al., 2016)	-	62.24	64.86	65.89
DRNN (Ouchi and Tsuboi, 2016)	-	63.28	66.70	68.41
SIRNN (Zhang et al., 2018)	-	72.59	77.13	78.53.
BERT (Devlin et al., 2019)	82.88	80.22	75.32	74.03
SA-BERT (Gu et al., 2020)	86.98	81.99	78.27	76.84
MPC-BERT (Gu et al., 2021)	89.54	84.21	80.67	78.98
BERT w/ GIFT	85.80 [†]	82.95 [†]	81.07 [†]	79.11 [†]
SA-BERT w/ GIFT	88.30 [†]	84.49 [†]	82.53 [†]	82.65 [†]
MPC-BERT w/ GIFT	90.18	85.85[†]	84.13[†]	83.61[†]

Results: Addressee Recognition

GIFT improves **BERT** by margins of **2.92%, 2.73%, 5.75% and 5.08%** on these test sets respectively in terms of **Precision (P@1)**

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
Preceding (Le et al., 2019)	-	55.73	55.63	55.62
SRNN (Ouchi and Tsuboi, 2016)	-	60.26	60.66	60.98
SHRNN (Serban et al., 2016)	-	62.24	64.86	65.89
DRNN (Ouchi and Tsuboi, 2016)	-	63.28	66.70	68.41
SIRNN (Zhang et al., 2018)	-	72.59	77.13	78.53.
BERT (Devlin et al., 2019)	82.88	80.22	75.32	74.03
SA-BERT (Gu et al., 2020)	86.98	81.99	78.27	76.84
MPC-BERT (Gu et al., 2021)	89.54	84.21	80.67	78.98
BERT w/ GIFT	85.80 [†]	82.95 [†]	81.07 [†]	79.11 [†]
SA-BERT w/ GIFT	88.30 [†]	84.49 [†]	82.53 [†]	82.65 [†]
MPC-BERT w/ GIFT	90.18	85.85[†]	84.13[†]	83.61[†]

Results: Addressee Recognition

GIFT improves **BERT** by margins of **2.92%, 2.73%, 5.75% and 5.08%** on these test sets respectively in terms of **Precision (P@1)**

improves **SA-BERT**
by margins of
**1.32%, 2.50%,
4.26% and 5.22%**
respectively

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
Preceding (Le et al., 2019)	-	55.73	55.63	55.62
SRNN (Ouchi and Tsuboi, 2016)	-	60.26	60.66	60.98
SHRNN (Serban et al., 2016)	-	62.24	64.86	65.89
DRNN (Ouchi and Tsuboi, 2016)	-	63.28	66.70	68.41
SIRNN (Zhang et al., 2018)	-	72.59	77.13	78.53.
BERT (Devlin et al., 2019)	82.88	80.22	75.32	74.03
SA-BERT (Gu et al., 2020)	86.98	81.99	78.27	76.84
MPC-BERT (Gu et al., 2021)	89.54	84.21	80.67	78.98
BERT w/ GIFT	85.80 [†]	82.95 [†]	81.07 [†]	79.11 [†]
SA-BERT w/ GIFT	88.30 [†]	84.49 [†]	82.53 [†]	82.65 [†]
MPC-BERT w/ GIFT	90.18	85.85[†]	84.13[†]	83.61[†]

Results: Addressee Recognition

GIFT improves **BERT** by margins of **2.92%, 2.73%, 5.75%** and **5.08%** on these test sets respectively in terms of **Precision (P@1)**

improves **SA-BERT**
by margins of
**1.32%, 2.50%,
4.26% and 5.22%**
respectively

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
Preceding (Le et al., 2019)	-	55.73	55.63	55.62
SRNN (Ouchi and Tsuboi, 2016)	-	60.26	60.66	60.98
SHRNN (Serban et al., 2016)	-	62.24	64.86	65.89
DRNN (Ouchi and Tsuboi, 2016)	-	63.28	66.70	68.41
SIRNN (Zhang et al., 2018)	-	72.59	77.13	78.53.
BERT (Devlin et al., 2019)	82.88	80.22	75.32	74.03
SA-BERT (Gu et al., 2020)	86.98	81.99	78.27	76.84
MPC-BERT (Gu et al., 2021)	89.54	84.21	80.67	78.98
BERT w/ GIFT	85.80 [†]	82.95 [†]	81.07 [†]	79.11 [†]
SA-BERT w/ GIFT	88.30 [†]	84.49 [†]	82.53 [†]	82.65 [†]
MPC-BERT w/ GIFT	90.18	85.85[†]	84.13[†]	83.61[†]

improves **MPC-BERT**
by margins of
**0.64%, 1.64%,
3.46% and 4.63%**
respectively

Results: Speaker Identification

GIFT improves **BERT** by margins of **13.71%, 27.50%, 29.14% and 28.82%** on these test sets respectively in terms of **P@1**

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
BERT (Devlin et al., 2019)	71.81	62.24	53.17	51.58
SA-BERT (Gu et al., 2020)	75.88	64.96	57.62	54.28
MPC-BERT (Gu et al., 2021)	83.54	67.56	61.00	58.52
BERT w/ GIFT	85.52 [†]	89.74 [†]	82.31 [†]	80.40 [†]
SA-BERT w/ GIFT	88.02 [†]	90.01 [†]	82.76 [†]	80.87 [†]
MPC-BERT w/ GIFT	90.50[†]	90.61[†]	84.12[†]	81.51[†]

Surprisingly effective for speaker identification!

Results: Speaker Identification

GIFT improves **BERT** by margins of **13.71%, 27.50%, 29.14% and 28.82%** on these test sets respectively in terms of **P@1**

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
BERT (Devlin et al., 2019)	71.81	62.24	53.17	51.58
SA-BERT (Gu et al., 2020)	75.88	64.96	57.62	54.28
MPC-BERT (Gu et al., 2021)	83.54	67.56	61.00	58.52
BERT w/ GIFT	85.52 [†]	89.74 [†]	82.31 [†]	80.40 [†]
SA-BERT w/ GIFT	88.02 [†]	90.01 [†]	82.76 [†]	80.87 [†]
MPC-BERT w/ GIFT	90.50[†]	90.61[†]	84.12[†]	81.51[†]

Surprisingly effective for speaker identification!

Results: Speaker Identification

GIFT improves **BERT** by margins of **13.71%, 27.50%, 29.14%** and **28.82%** on these test sets respectively in terms of **P@1**

improves **SA-BERT**
by margins of
12.14%, 25.05%,
25.14% and
26.59%
respectively

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
BERT (Devlin et al., 2019)	71.81	62.24	53.17	51.58
SA-BERT (Gu et al., 2020)	75.88	64.96	57.62	54.28
MPC-BERT (Gu et al., 2021)	83.54	67.56	61.00	58.52
BERT w/ GIFT	85.52 [†]	89.74 [†]	82.31 [†]	80.40 [†]
SA-BERT w/ GIFT	88.02 [†]	90.01 [†]	82.76 [†]	80.87 [†]
MPC-BERT w/ GIFT	90.50[†]	90.61[†]	84.12[†]	81.51[†]

Surprisingly effective for speaker identification!

Results: Speaker Identification

GIFT improves **BERT** by margins of **13.71%, 27.50%, 29.14%** and **28.82%** on these test sets respectively in terms of **P@1**

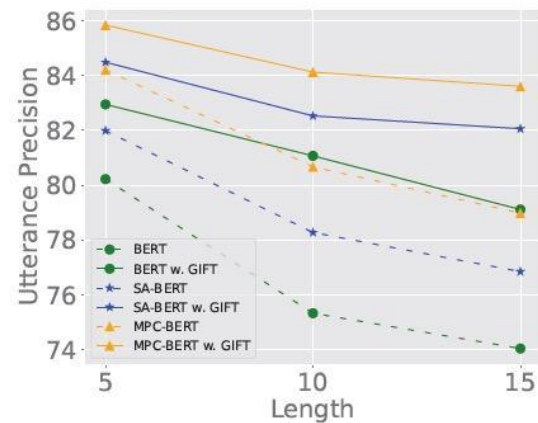
improves **SA-BERT**
by margins of
12.14%, 25.05%,
25.14% and
26.59%
respectively

	Hu et al. (2019)	Ouchi and Tsuboi (2016)		
		Len-5	Len-10	Len-15
BERT (Devlin et al., 2019)	71.81	62.24	53.17	51.58
SA-BERT (Gu et al., 2020)	75.88	64.96	57.62	54.28
MPC-BERT (Gu et al., 2021)	83.54	67.56	61.00	58.52
BERT w/ GIFT	85.52 [†]	89.74 [†]	82.31 [†]	80.40 [†]
SA-BERT w/ GIFT	88.02 [†]	90.01 [†]	82.76 [†]	80.87 [†]
MPC-BERT w/ GIFT	90.50[†]	90.61[†]	84.12[†]	81.51[†]

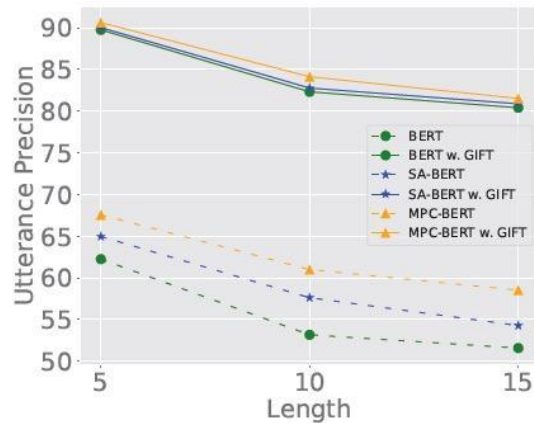
improves
MPC-BERT by
margins of **6.96%,**
23.05%, 23.12%
and **22.99%**
respectively

Surprisingly effective for speaker identification!

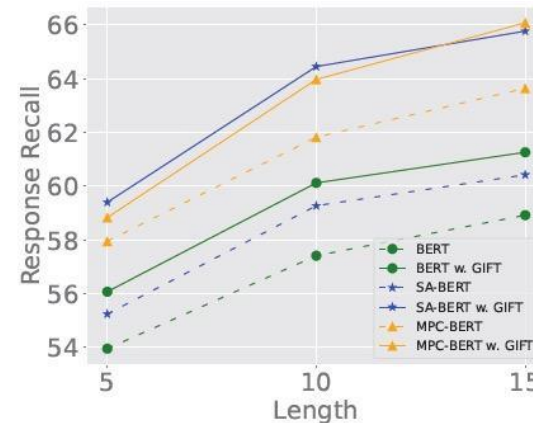
Performance Change at Different Lengths



(a) Addressee Recognition



(b) Speaker Identification

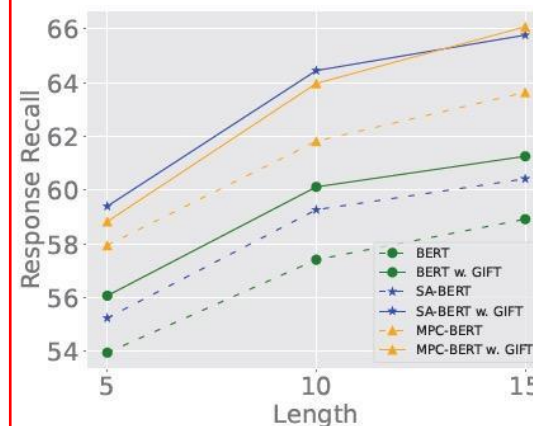
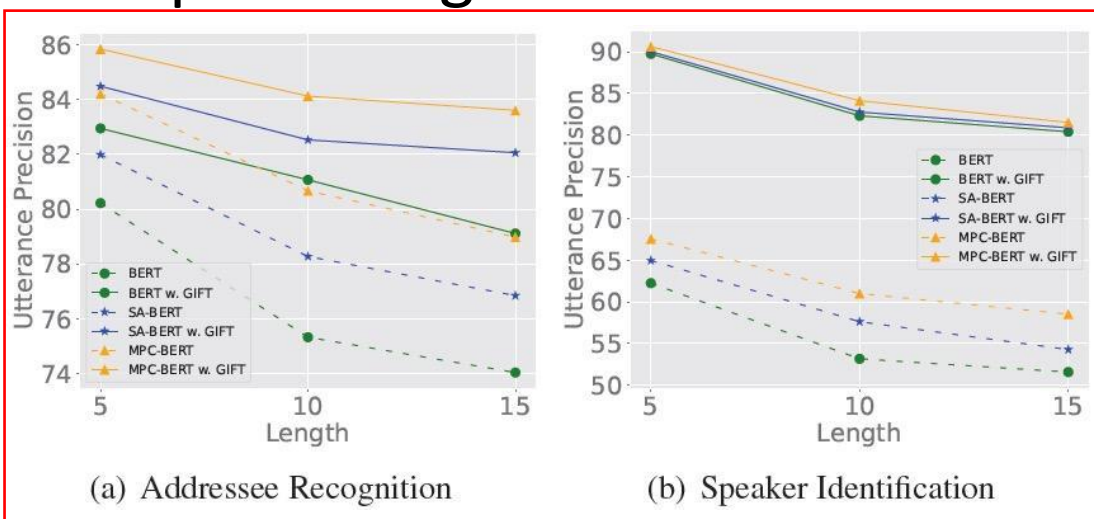


(c) Response Selection

Performance Change at Different Lengths

Results: the performance of **addressee recognition** and **speaker identification** **dropped** as the session length increased

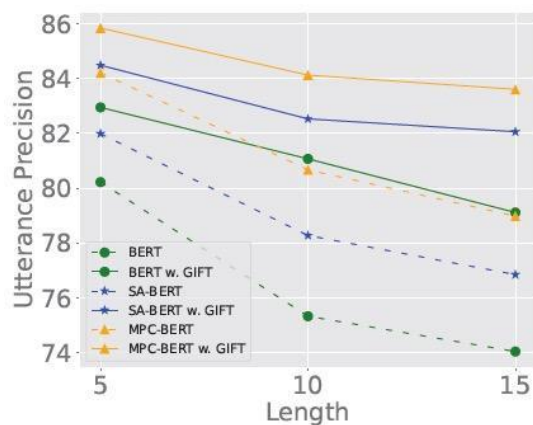
Analysis: **longer** sessions always contain **more interlocutors** which increase the difficulties of predicting interlocutors



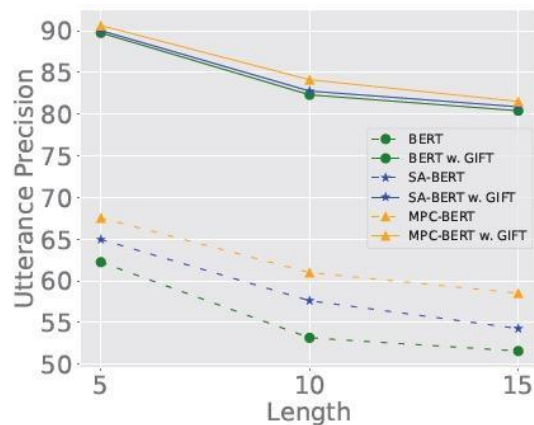
Performance Change at Different Lengths

Results: the performance of **response selection** was significantly **improved** as the session length increased

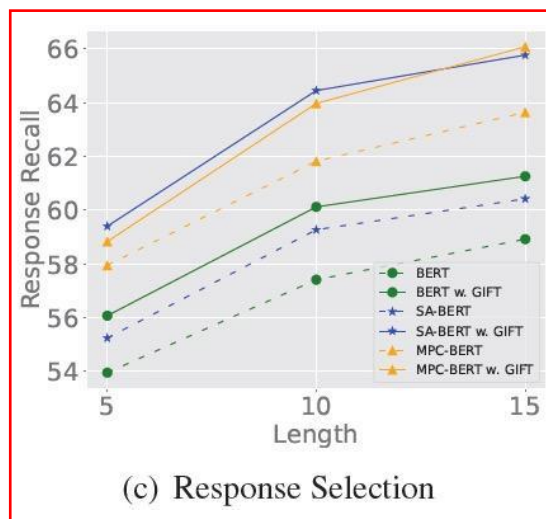
Analysis: **longer** sessions **enrich** the representations of contexts with more details which benefit response selection



(a) Addressee Recognition



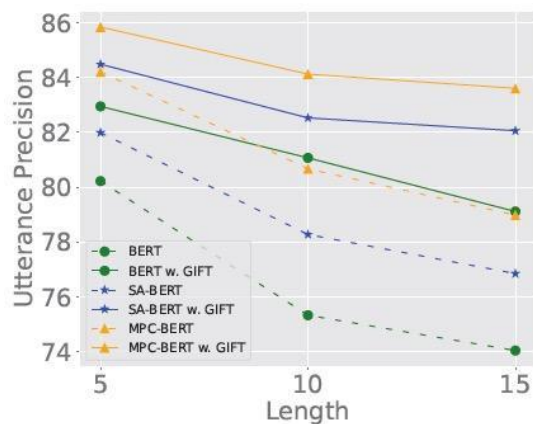
(b) Speaker Identification



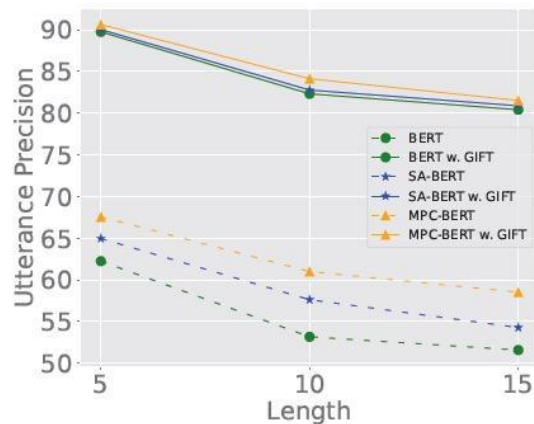
(c) Response Selection

Performance Change at Different Lengths

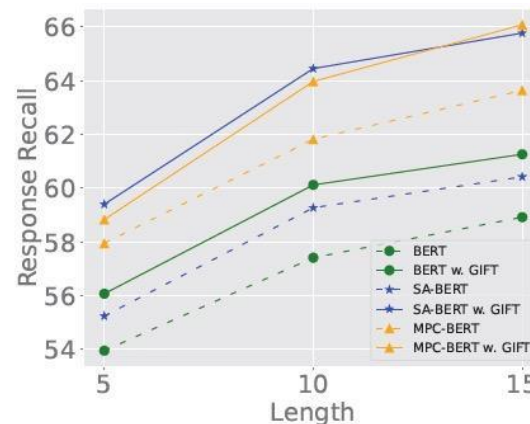
As the session length increased, the performance of **models with GIFT** dropped more slightly on addressee recognition and speaker identification, and **enlarged more** on response selection, than the **models without GIFT** in most **14 out of 18** cases (2 margins for 3 models on 3 tasks)



(a) Addressee Recognition



(b) Speaker Identification



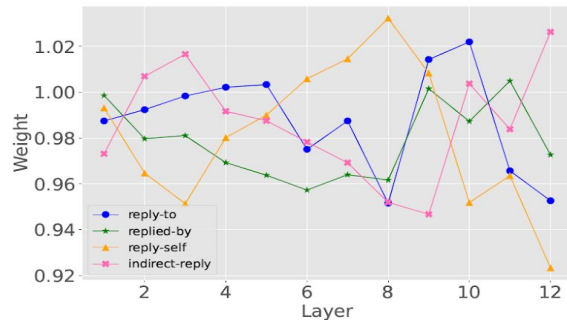
(c) Response Selection

	Len 5 → Len 10	Len 10 → Len 15
AR (P@1)		
BERT	-4.90	-1.29
BERT w. GIFT	-1.88 [‡]	-1.96
SA-BERT	-3.72	-1.43
SA-BERT w. GIFT	-1.96 [‡]	-0.47 [‡]
MPC-BERT	-3.54	-1.69
MPC-BERT w. GIFT	-1.72 [‡]	-0.52 [‡]
SI (P@1)		
BERT	-9.07	-1.59
BERT w. GIFT	-7.43 [‡]	-1.91
SA-BERT	-7.34	-3.34
SA-BERT w. GIFT	-7.25 [‡]	-1.89 [‡]
MPC-BERT	-6.56	-2.48
MPC-BERT w. GIFT	-6.49 [‡]	-2.61
RS (R ₁₀ @1)		
BERT	+3.46	+1.51
BERT w. GIFT	+4.05 [‡]	+1.14
SA-BERT	+4.03	+1.15
SA-BERT w. GIFT	+5.05 [‡]	+1.32 [‡]
MPC-BERT	+3.87	+1.82
MPC-BERT w. GIFT	+5.14 [‡]	+2.11 [‡]

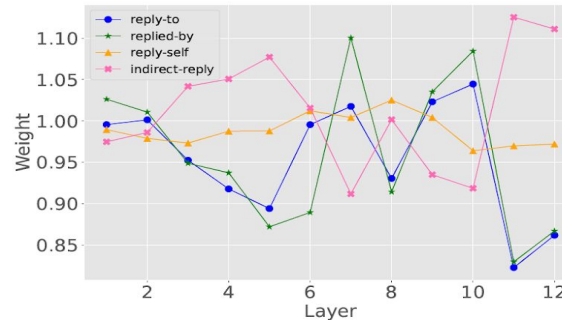
Table 6: Performance change of models as the session length increased on the test sets of Ouchi and Tsuboi (2016). For models with GIFT, numbers marked with [‡] denoted larger performance improvement or less performance drop compared with the corresponding models without GIFT.

Visualization of Weights

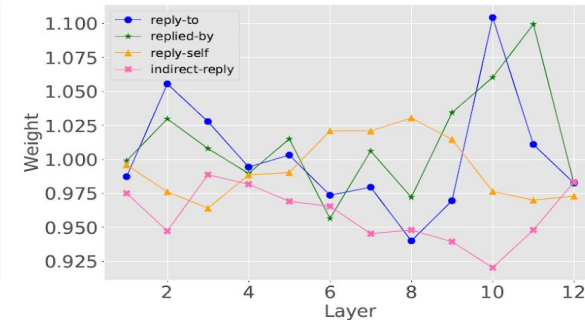
- The changing trends of **reply-to** and **replied-by** edges were **roughly the same**, while the values of these two edges were **always different**
- The values of the **indirect-reply** edge were always the **minimum at the beginning**, and surprisingly became the **maximum in the last layer**,
 - ✓ less attention to irrelevant utterances to themselves at first glance
 - ✓ After comprehending the most relevant utterances, turn to indirectly related ones in context for fully understanding the entire conversation



(a) Addressee Recognition



(b) Speaker Identification



(c) Response Selection

Figure 4: The weights of four types of edges in different encoding layers of MPC-BERT trained on [Hu et al. \(2019\)](#).

Q & A



We will be back at 15:30!

Section 4: Addressee Modeling (cont'd)

Representative Tasks

- Addressee recognition is tasked to directly recognize the addressee of target utterances given the interlocutor set in this conversation (explicit addressee modeling)
- **Dialogue disentanglement** aims at disentangling a whole conversation from a data stream into several threads via the **underlying reply relationships**, so that each thread is about a specific topic (**implicit addressee modeling**)

Dialogue Disentanglement

- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...@...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?.. this digital camera age kills me when I have to scale photos :s
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- [03:06] part of imagemagick
- === E-bola [...@...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] BurgerMann, i used that to convert 100's of photos in one command
- [03:06] Oh... I'll have a look.. thx =)



A whole conversation from
a data stream with multiple
threads interleaved

Dialogue Disentanglement

- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...@...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?.. this digital camera age kills me when I have to scale photos :s
- [03:06] delire, yes
- [03:06] BurgerMann, convert
- [03:06] part of imagemagick
- === E-bola [...@...] has left #ubuntu []
- [03:06] BurgerMann: ImageMagick
- [03:06] BurgerMann, i used that to convert 100's of photos in one command
- [03:06] Oh... I'll have a look.. thx =)



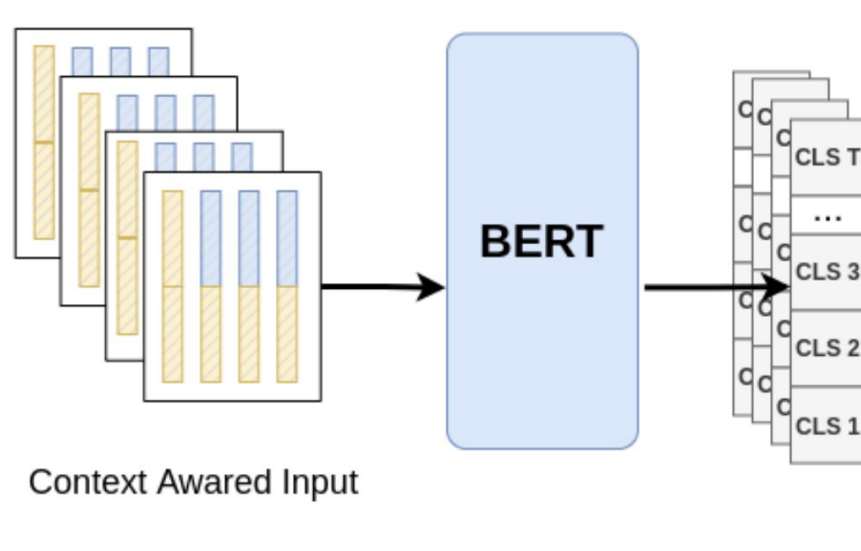
Easy to understand and
respond appropriately after
disentanglement

DialBERT

Tianda Li, et al. *DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement*. 2020.

Tianda Li, et al. *Conversation- and Tree-Structure Losses for Dialogue Disentanglement*. DialDoc 2022.

DialBERT

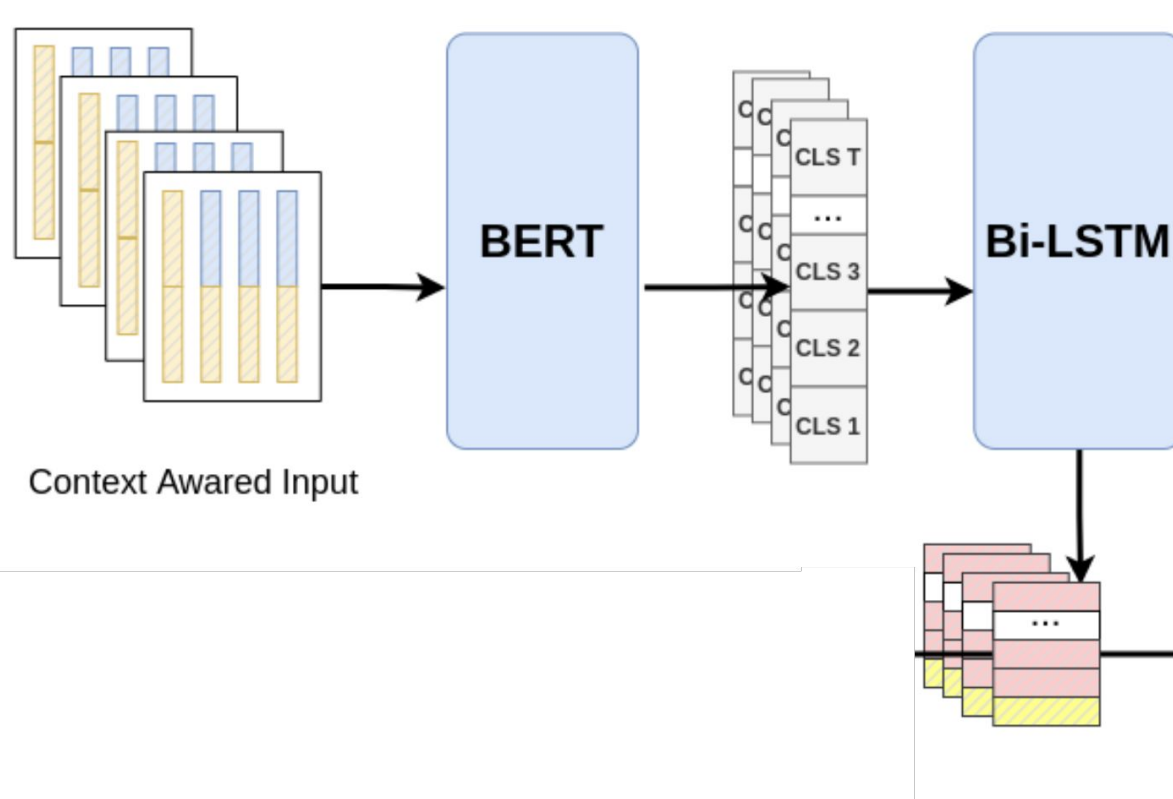


Capture **local** semantics by concatenating **target** with each of **context utterance**

Tianda Li, et al. *DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement*. 2020.

Tianda Li, et al. *Conversation- and Tree-Structure Losses for Dialogue Disentanglement*. DialDoc 2022.

DialBERT

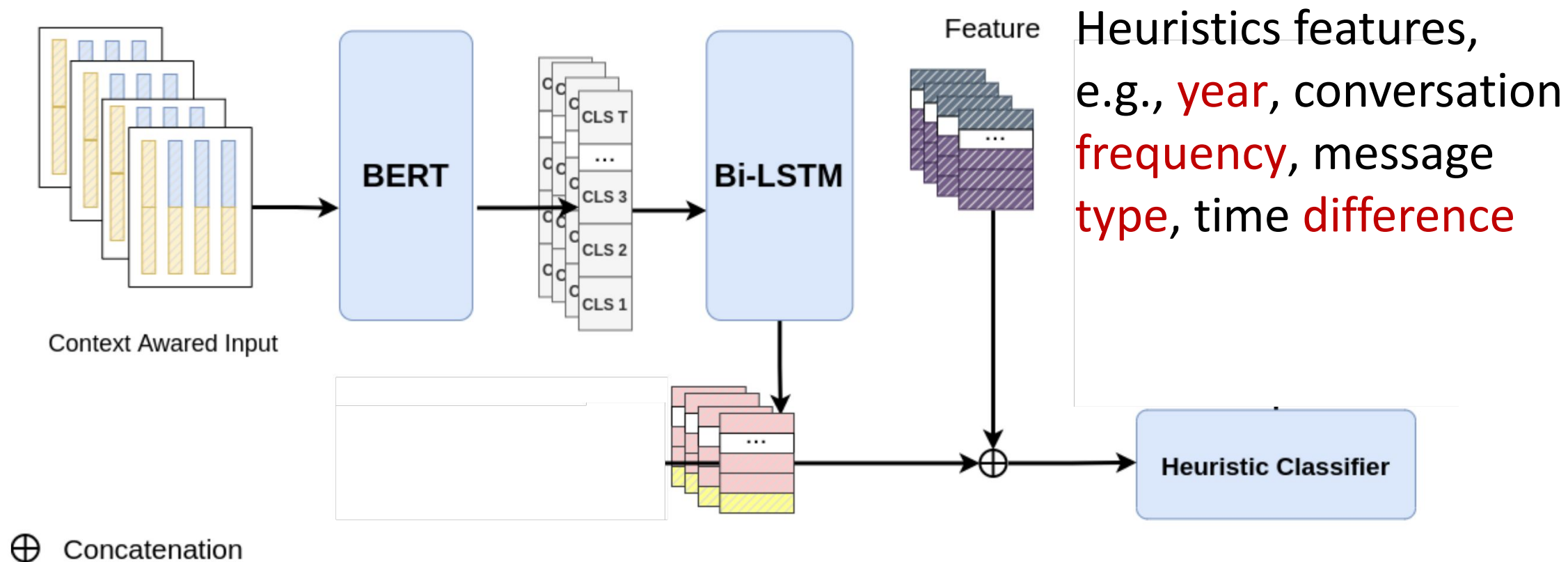


Capture **global** semantics across different message pairs to enhance context

Tianda Li, et al. *DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement*. 2020.

Tianda Li, et al. *Conversation- and Tree-Structure Losses for Dialogue Disentanglement*. DialDoc 2022.

DialBERT

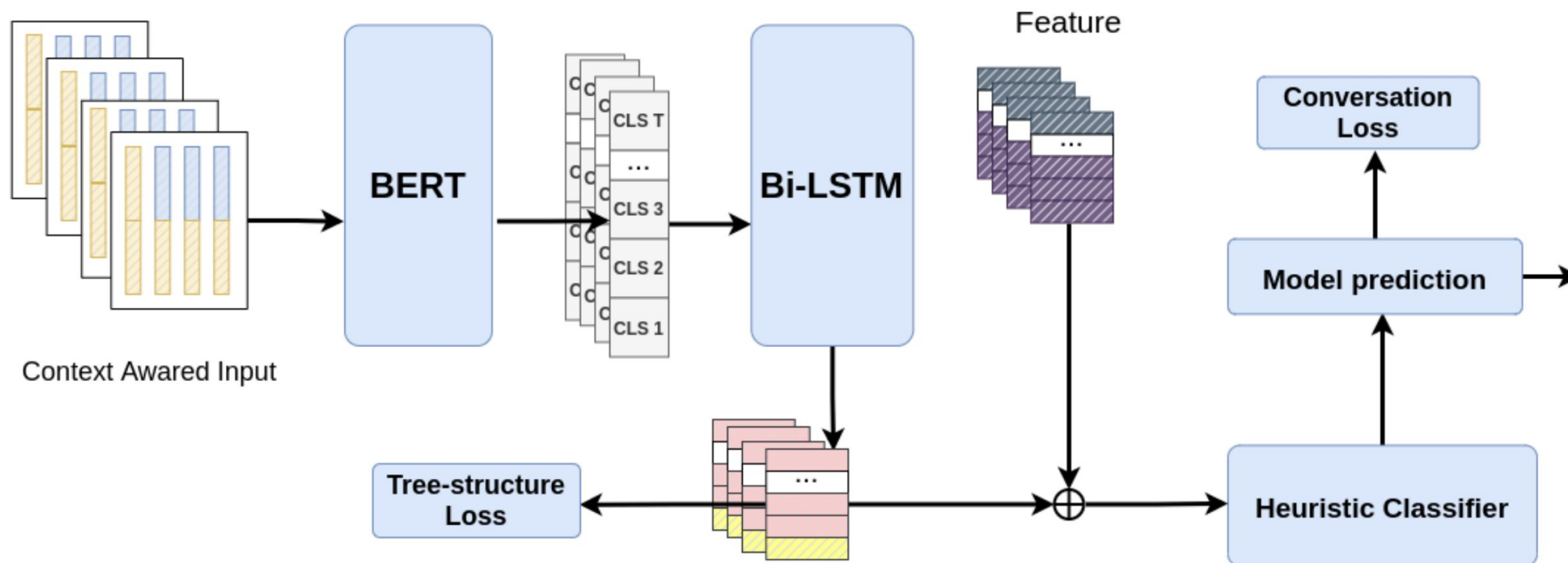


Tianda Li, et al. *DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement*. 2020.

Tianda Li, et al. *Conversation- and Tree-Structure Losses for Dialogue Disentanglement*. DialDoc 2022.

DialBERT

Distinguish **which conversation structure** the target message belong to



$\oplus$ Concatenation

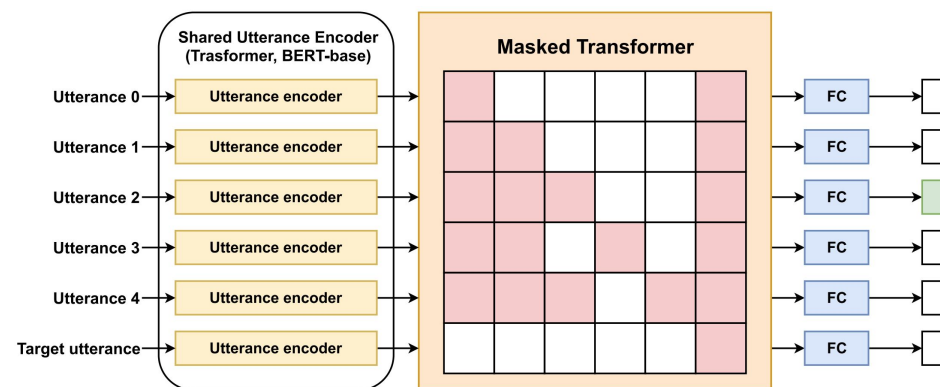
Further distinguish **the ancestor messages** of the target message in this structure

Tianda Li, et al. *DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement*. 2020.

Tianda Li, et al. *Conversation- and Tree-Structure Losses for Dialogue Disentanglement*. DialDoc 2022.

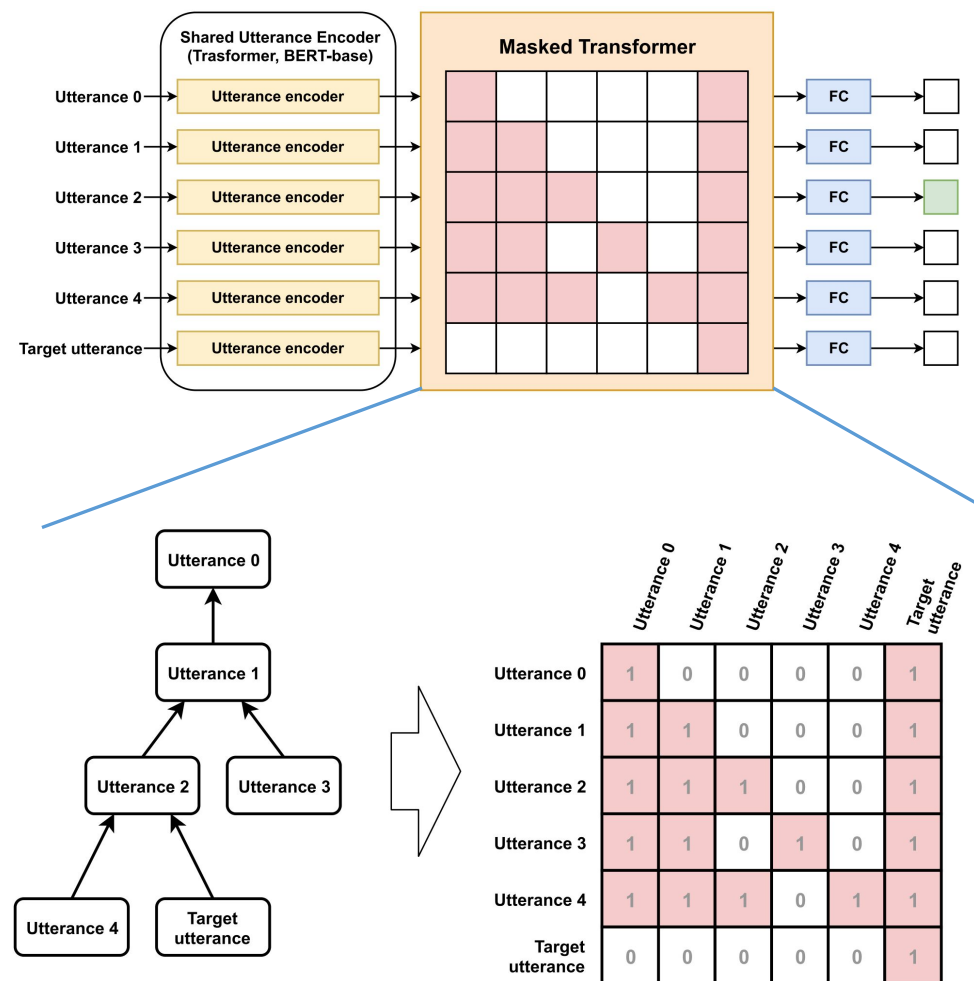
Masked Hierarchical Transformer

Learn conversation structures via **masking** to denote which history utterances are **attendable** to guide and aggregate the **ancestor flow**



Masked Hierarchical Transformer

Learn conversation structures via **masking** to denote which history utterances are **attendable** to guide and aggregate the **ancestor flow**

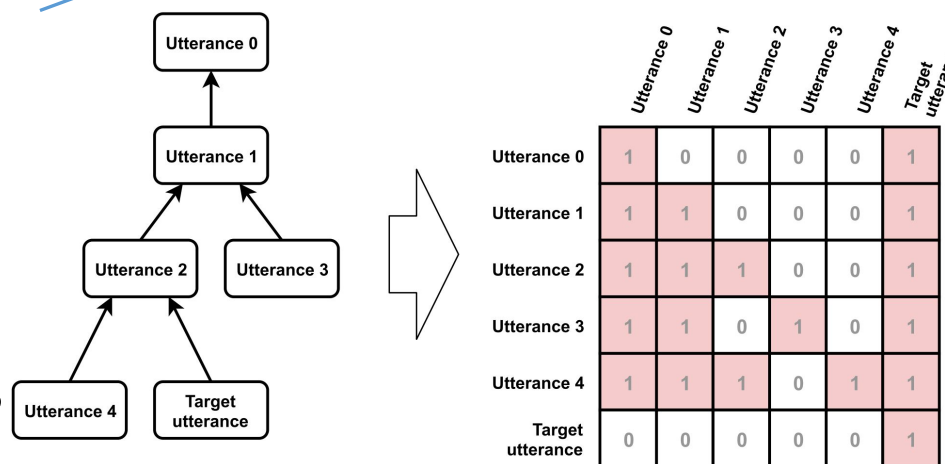
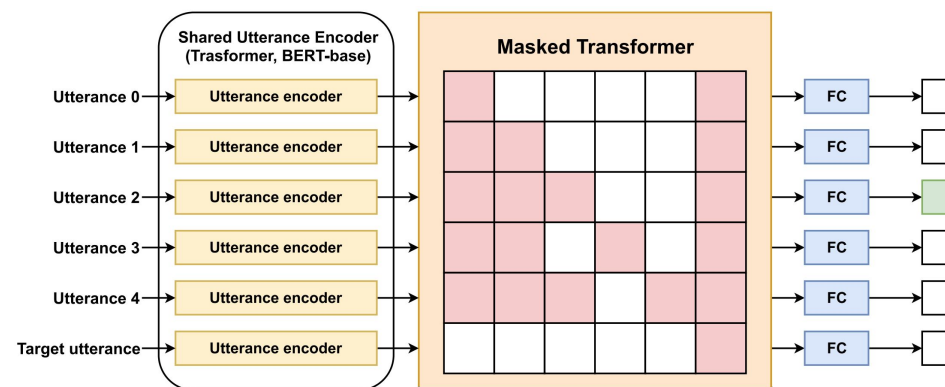


Masked Hierarchical Transformer

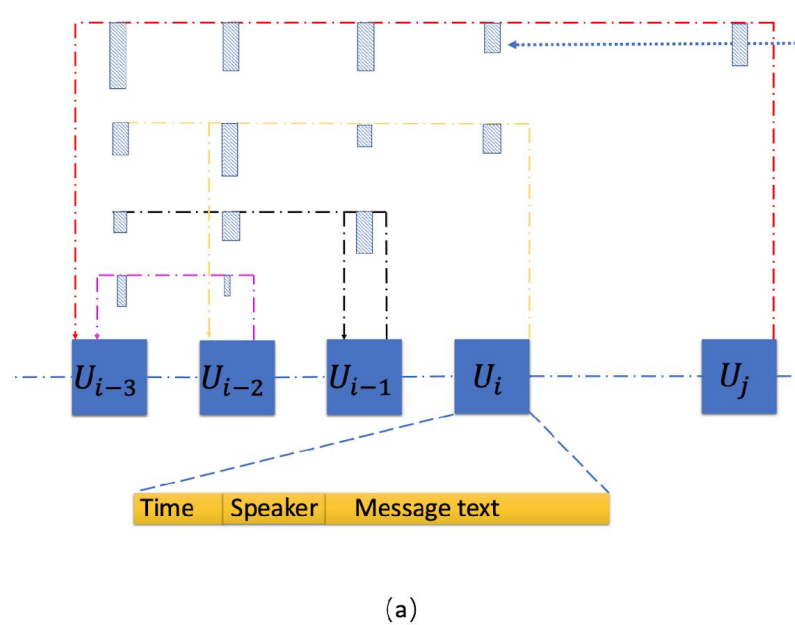
Learn conversation structures via **masking** to denote which history utterances are **attendable** to guide and aggregate the **ancestor flow**

Masking properties:

- Attend to target utterance
- Attend to itself
- Non-target utterances attend to its ancestors in the conversation graph
- Not attend to all remaining utterances



Pointer Network



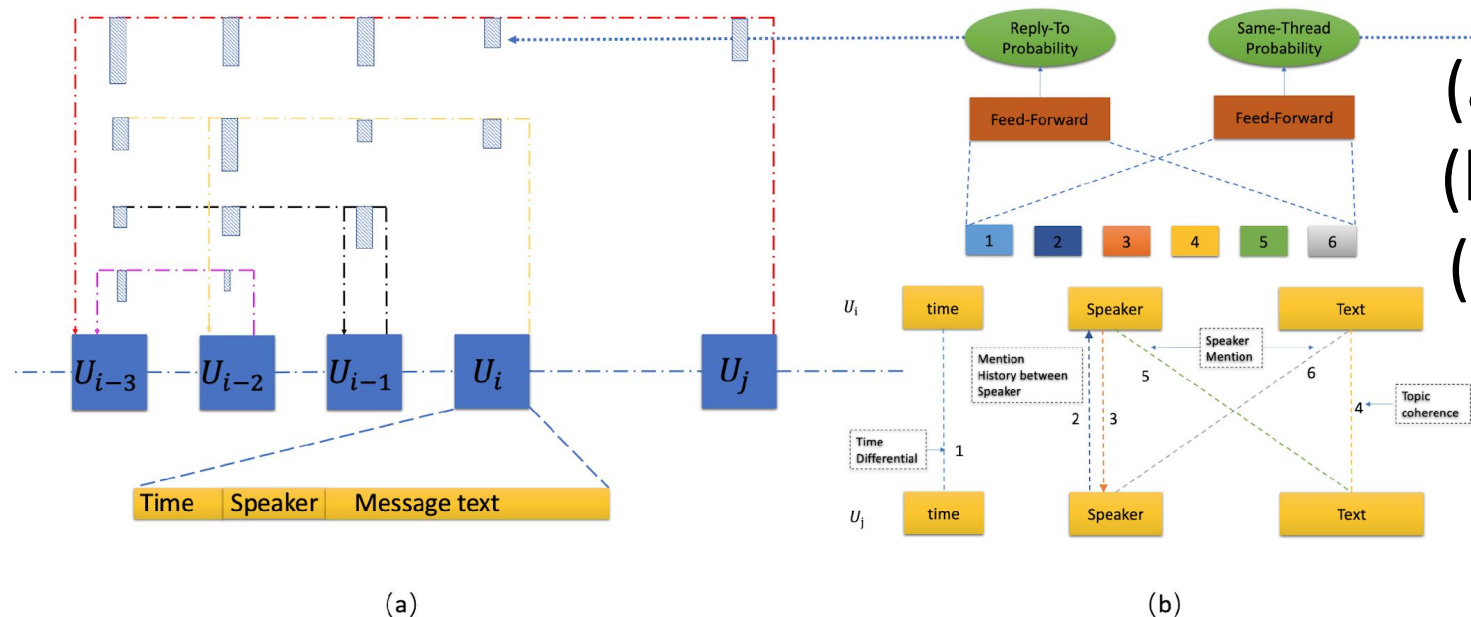
Encode each utterance
consisting of three parts: **time**,
speaker and **message text**

Pointer Network

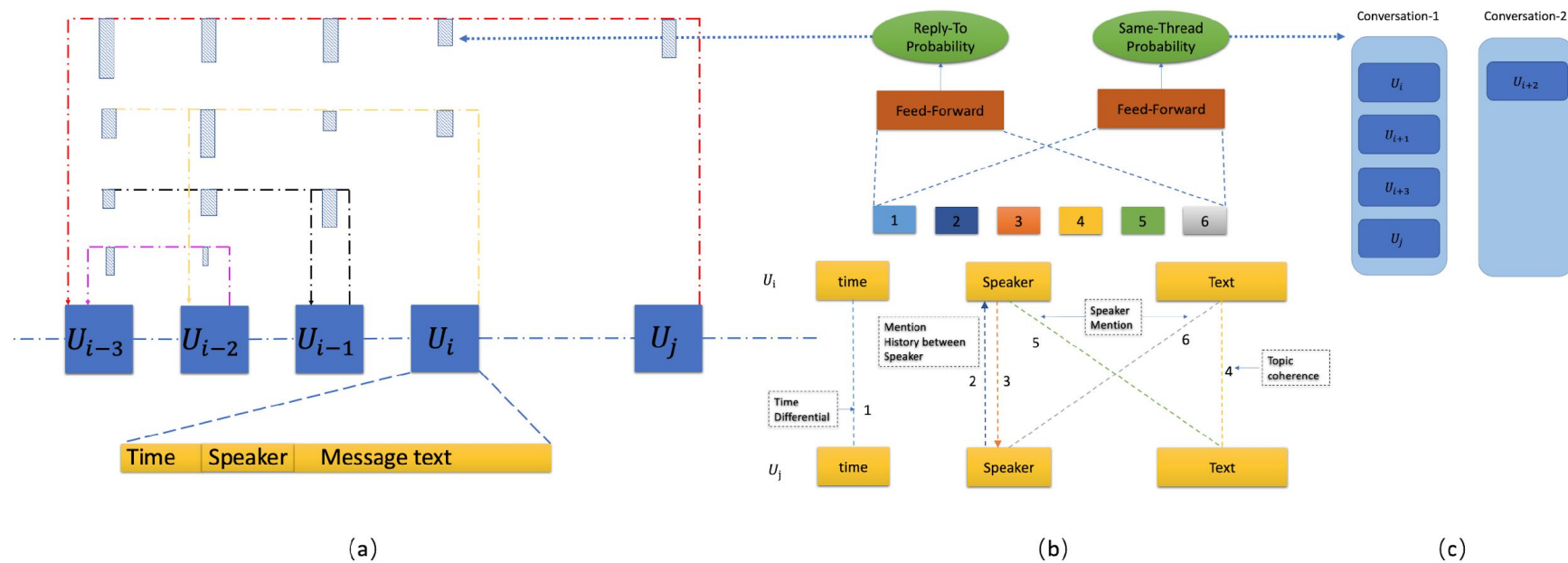
Capture **first-order inter-utterance interactions** through five features:

- (a) Time difference
- (b) Mention
- (c) Mention history
- (d) Topic coherence
- (e) Pointing

Model link prediction as a **multinomial distribution** over the previous utterances



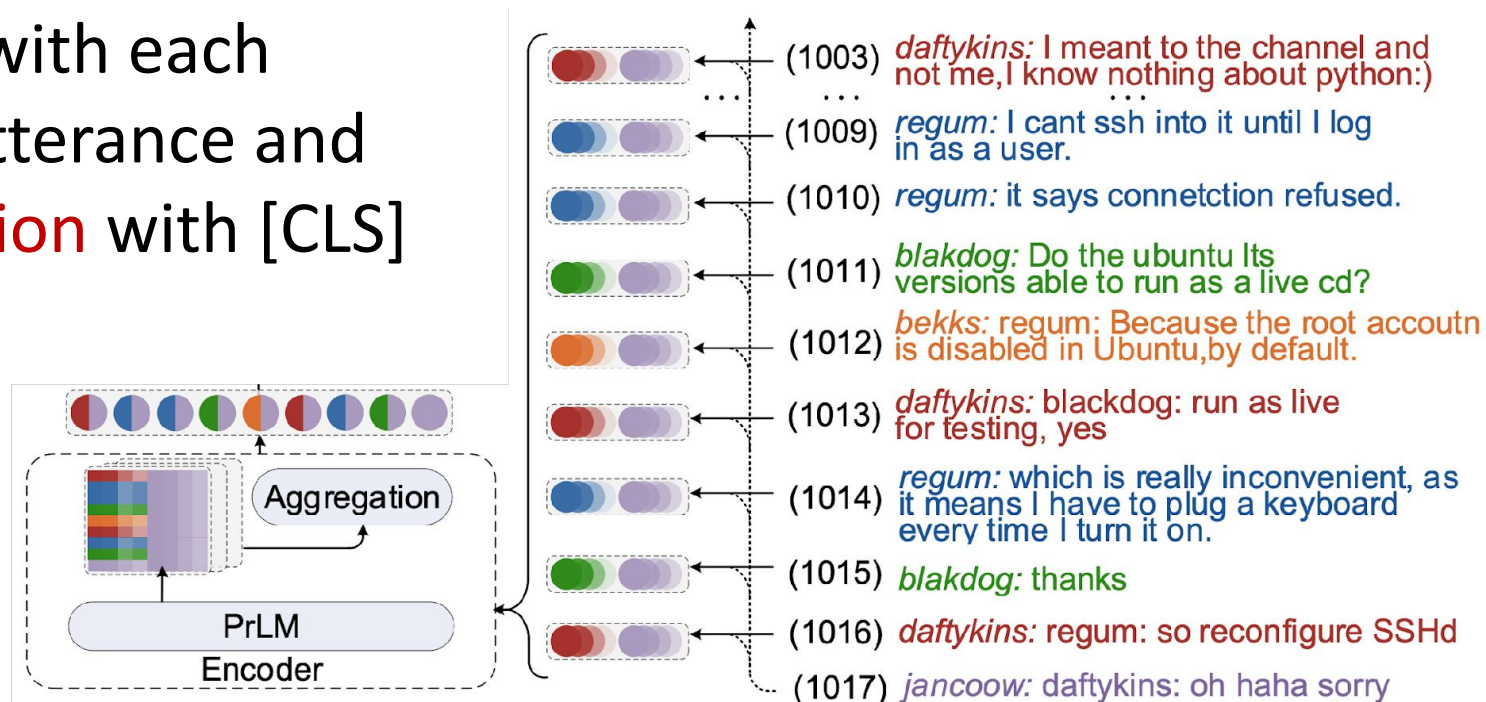
Pointer Network



Capture **higher-order** contextual information with pairwise classification
deciding **whether two utterances are in the same conversation**

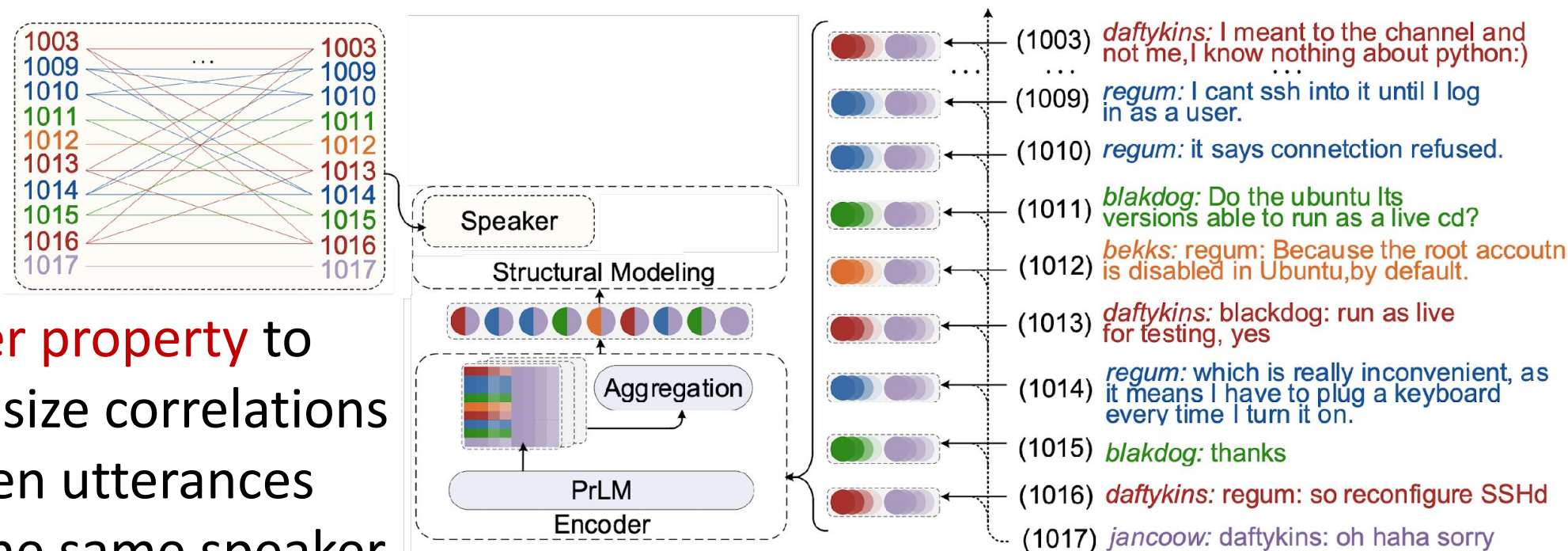
Structural Characterization

Pairwise encoding with each parent candidate utterance and **Utterance aggregation** with [CLS] representation

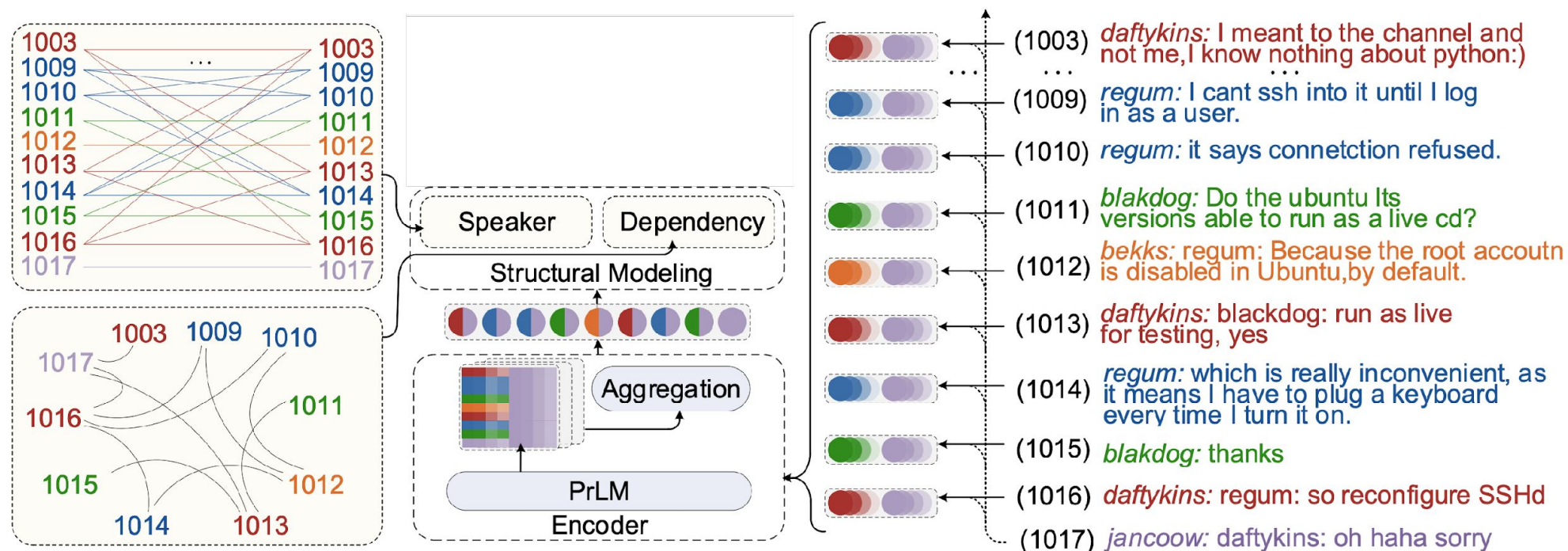


Structural Characterization

Speaker property to emphasize correlations between utterances from the same speaker via attention mask

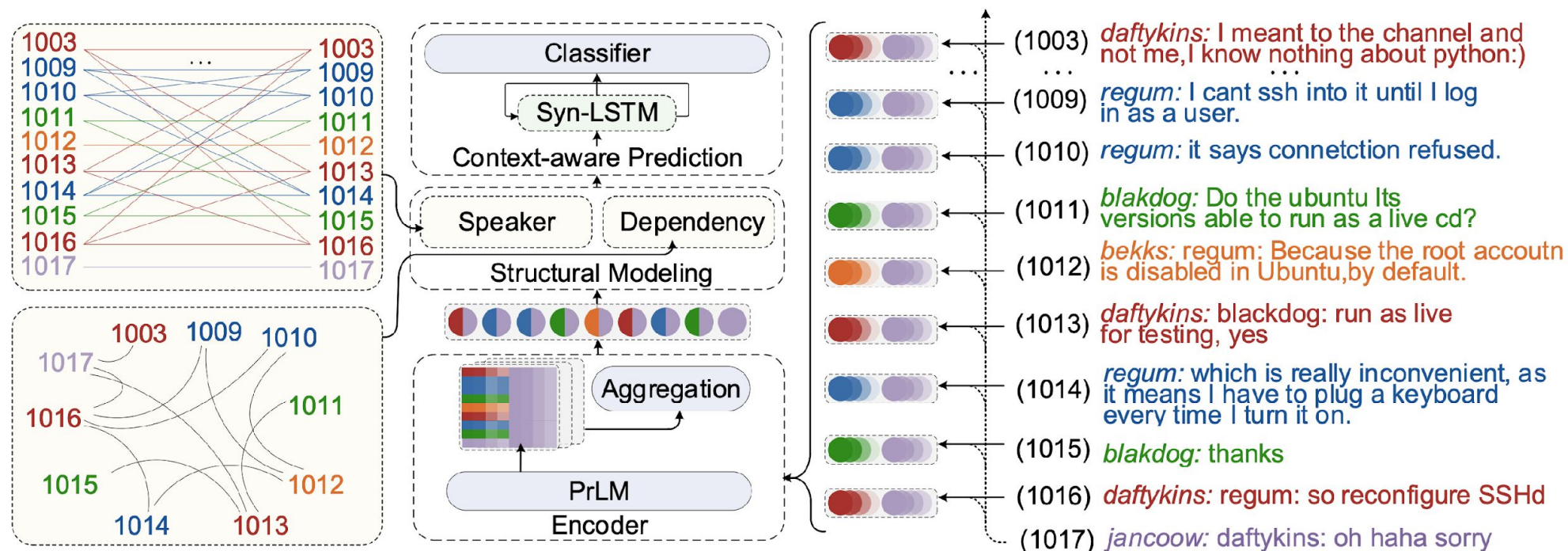


Structural Characterization



Enhance **reference dependency** via relational GCN
utterance as a vertex and “reply to” as an edge

Structural Characterization



Fuse **structure-aware** and **original** representation via Syn-LSTM for contextualization

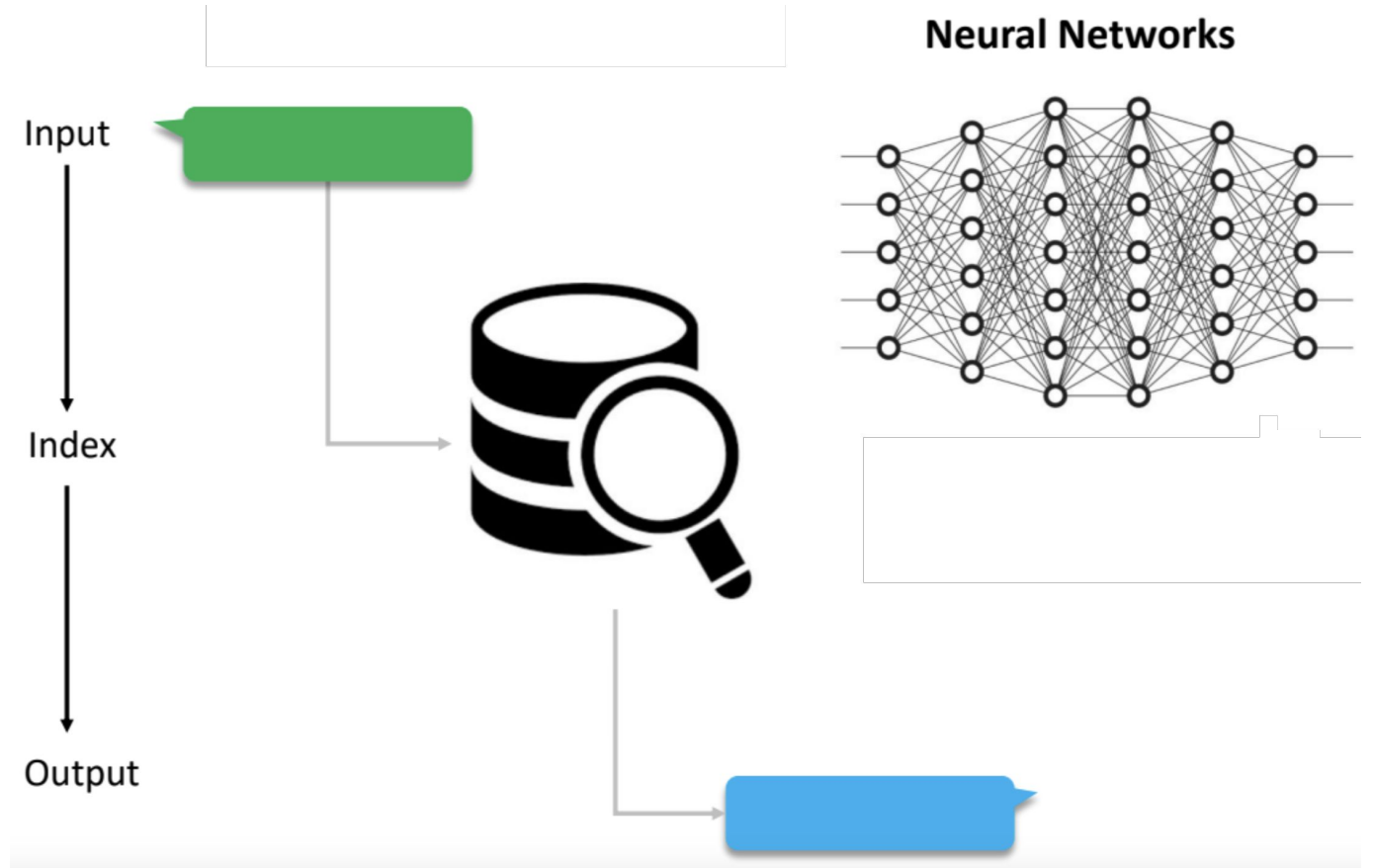
Section 5: Response Modeling

Representative Tasks

- **Response selection** aims at **selecting** the best-matched response from a set of candidates, given the context of a multi-turn conversation (**retrieval**-based)
- Response generation synthesizes a response with a natural language generative model by maximizing its generation probability given the previous conversation history (generation-based)

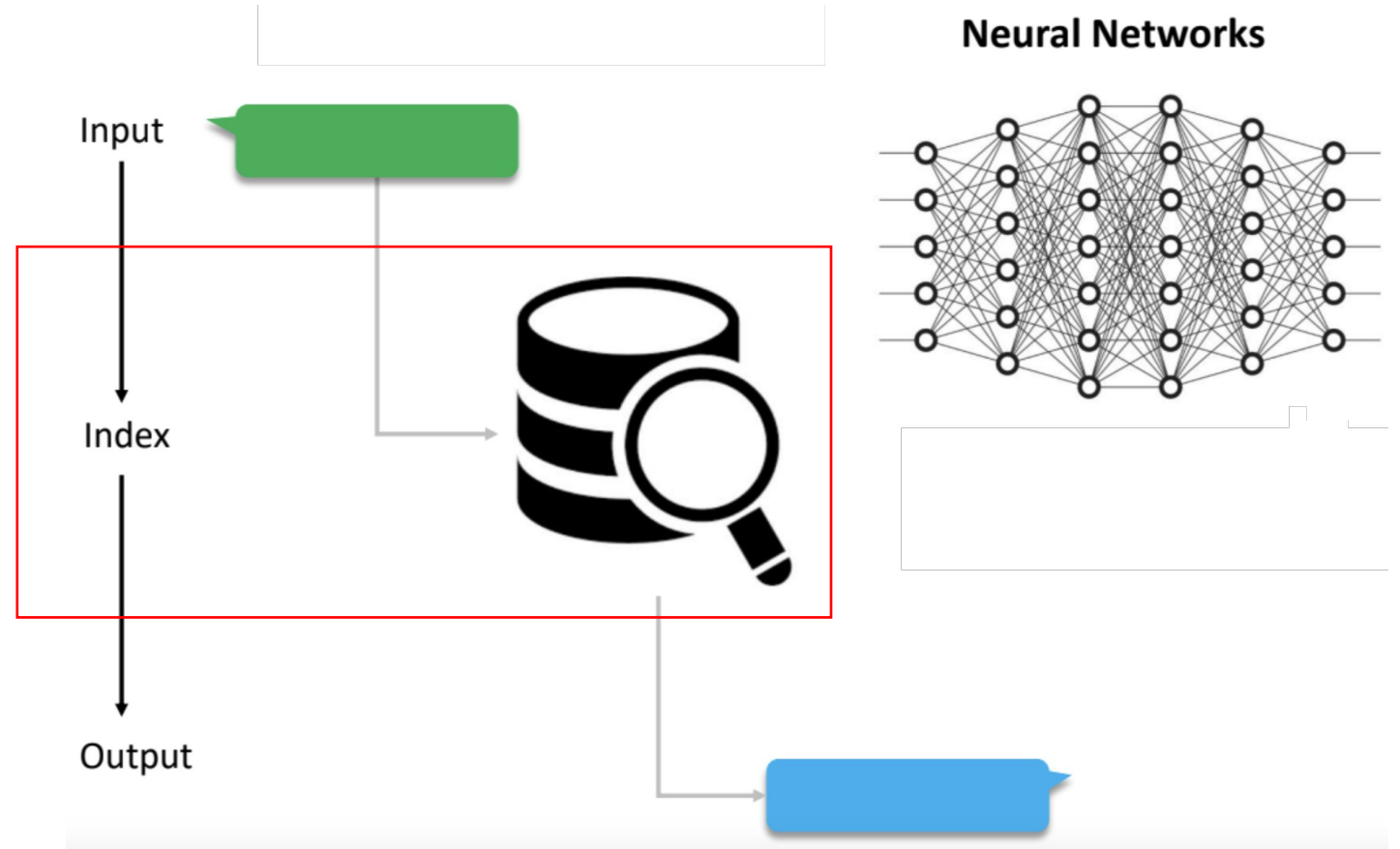
Response Selection

Not only:
Semantics
Consistency
Interactiveness



Response Selection

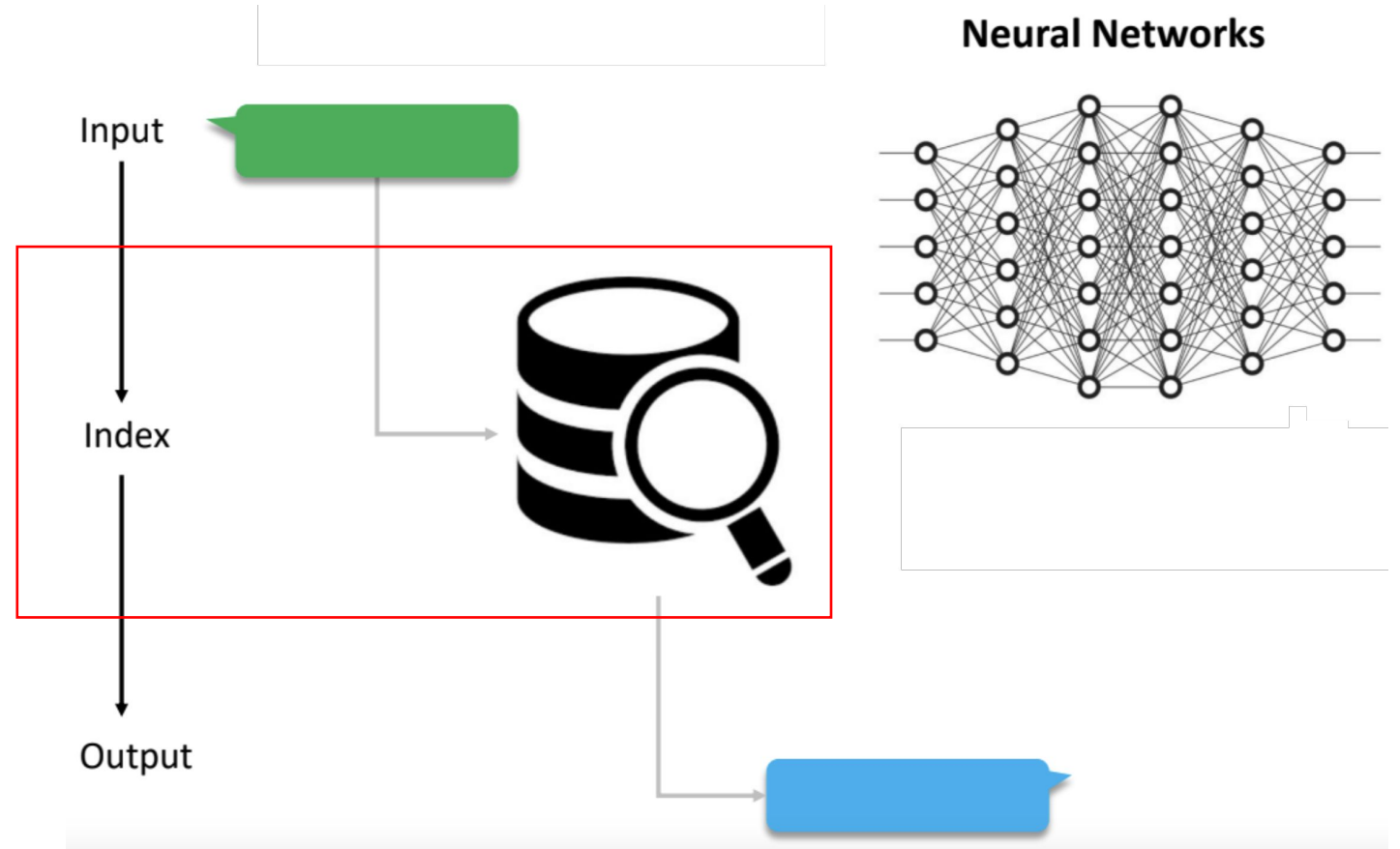
Not only:
Semantics
Consistency
Interactiveness



Response Selection

Not only:

- Semantics
- Consistency
- Interactiveness



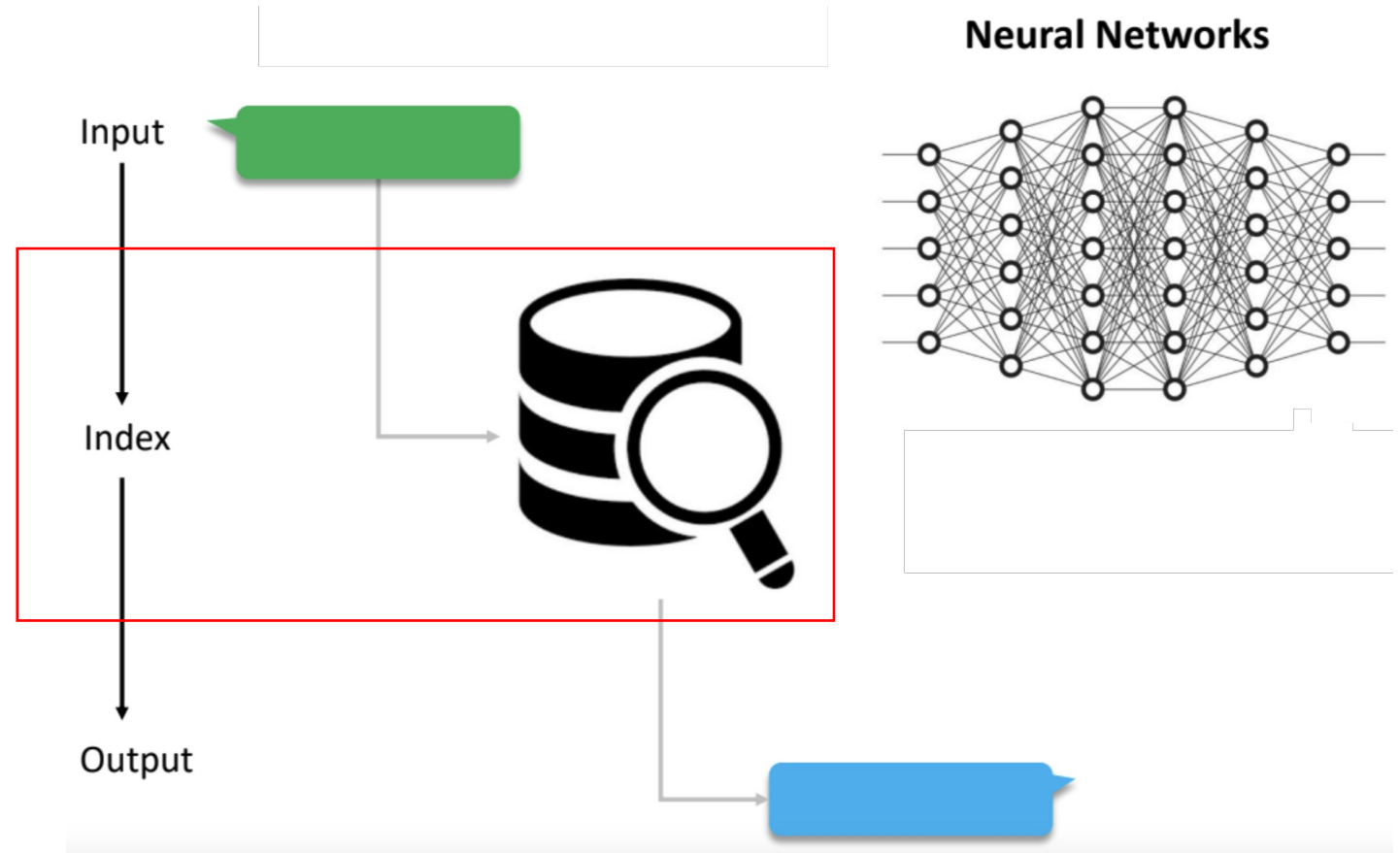
Response Selection

Not only:

- Semantics
- Consistency
- Interactiveness

But also:

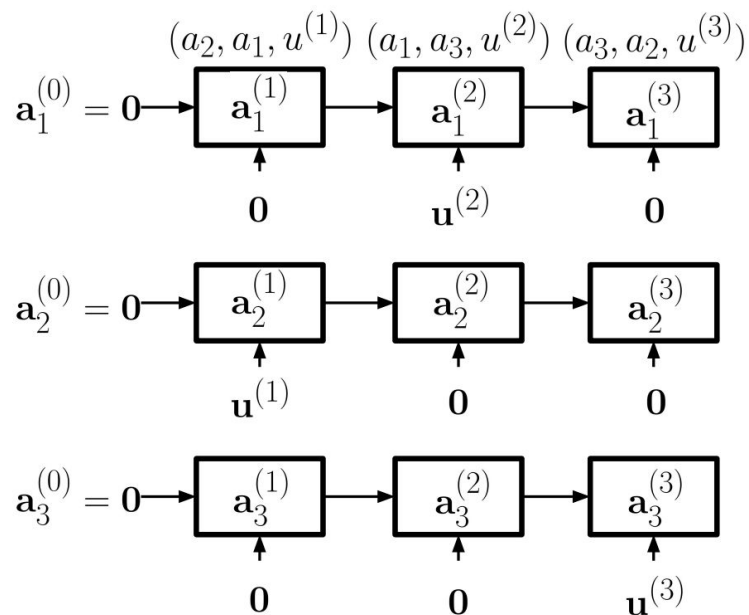
- Conversation structure
- Topic transition



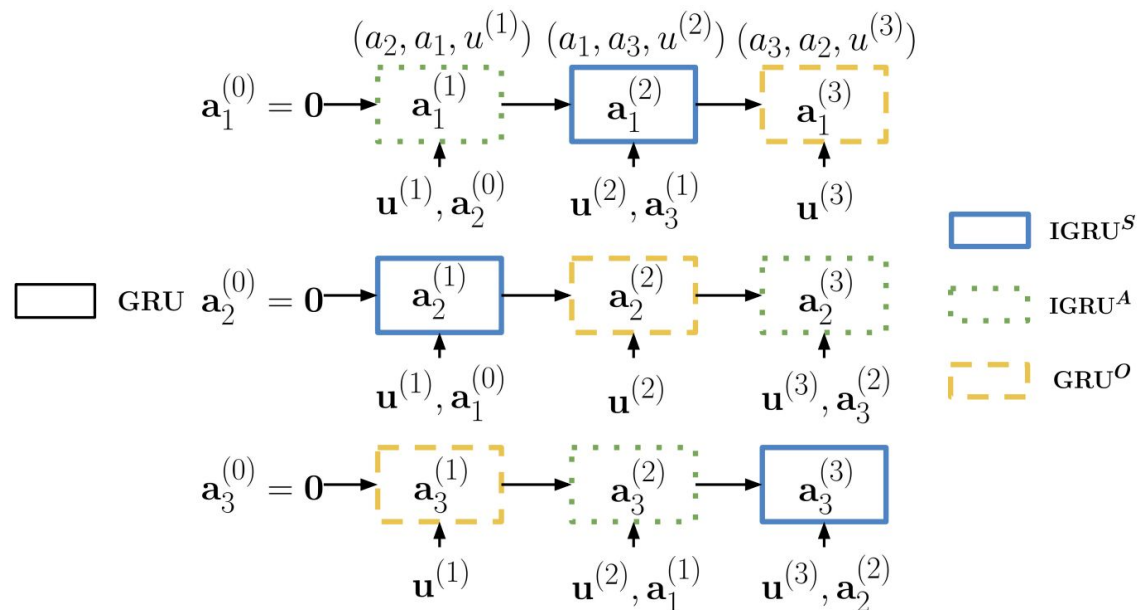
DRNN & SIRNN

Jointly model interlocutors and utterances

But **not fuse users' states** into the utterance embeddings



DRNN



SIRNN

Ouchi and Tsuboi. *Addressee and Response Selection for Multi-Party Conversation*. EMNLP 2016.

Topic-BERT

Frame response selection as **dynamic topic tracking**

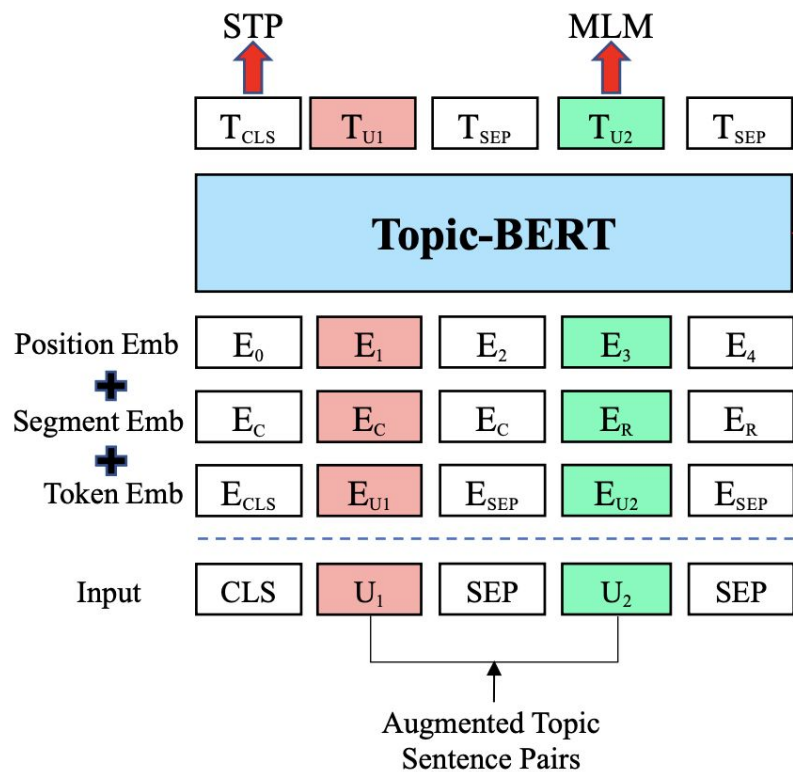
Topic-BERT

Frame response selection as **dynamic topic tracking**

- remain the **same topic** as going from **context** to **response**

Topic-BERT

Frame response selection as **dynamic topic tracking**



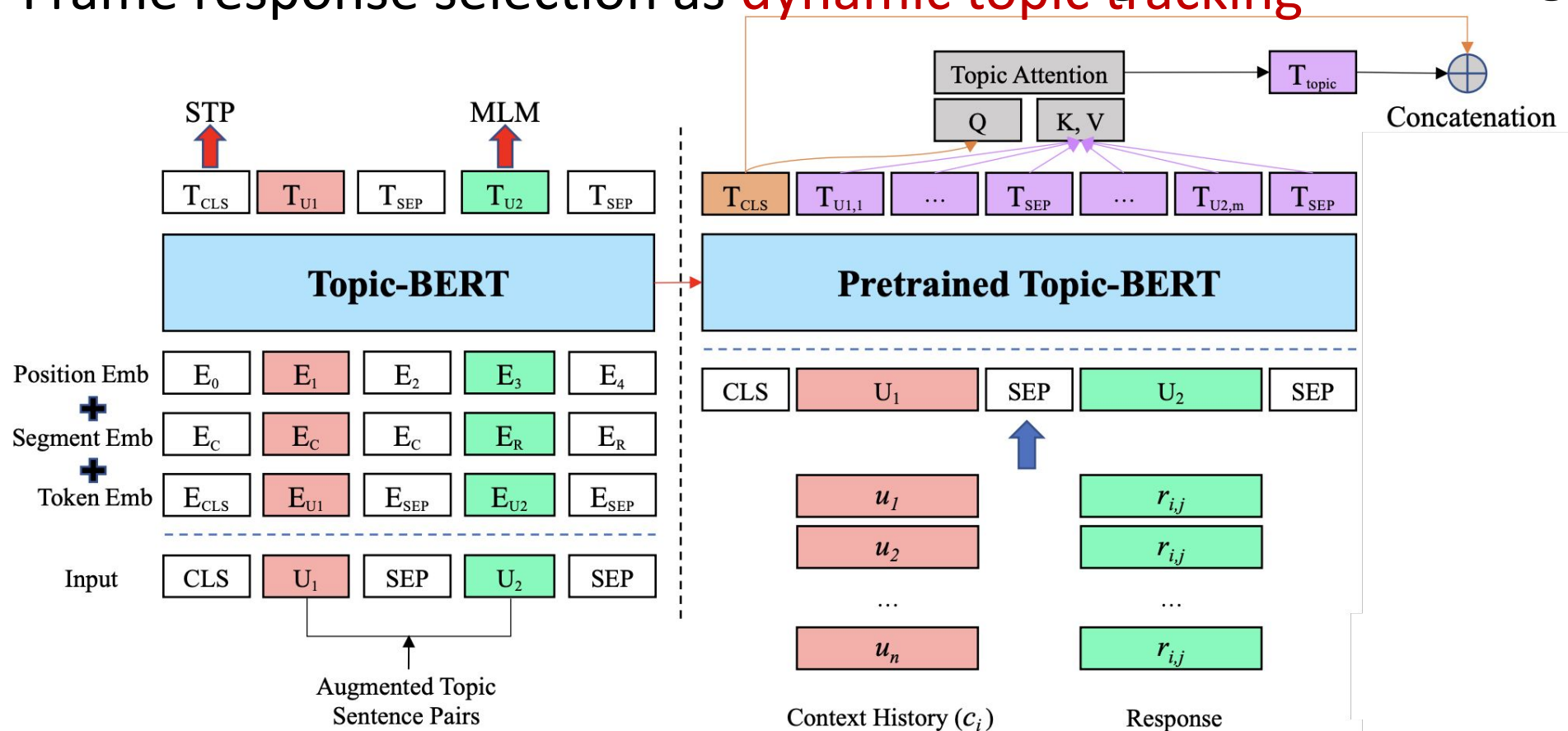
Pretrain with

- a) **Same Topic Prediction (STP)**: if a pair of utterance in a single-threaded conversation
- b) **Masked Language Modeling (MLM)**

Topic-BERT

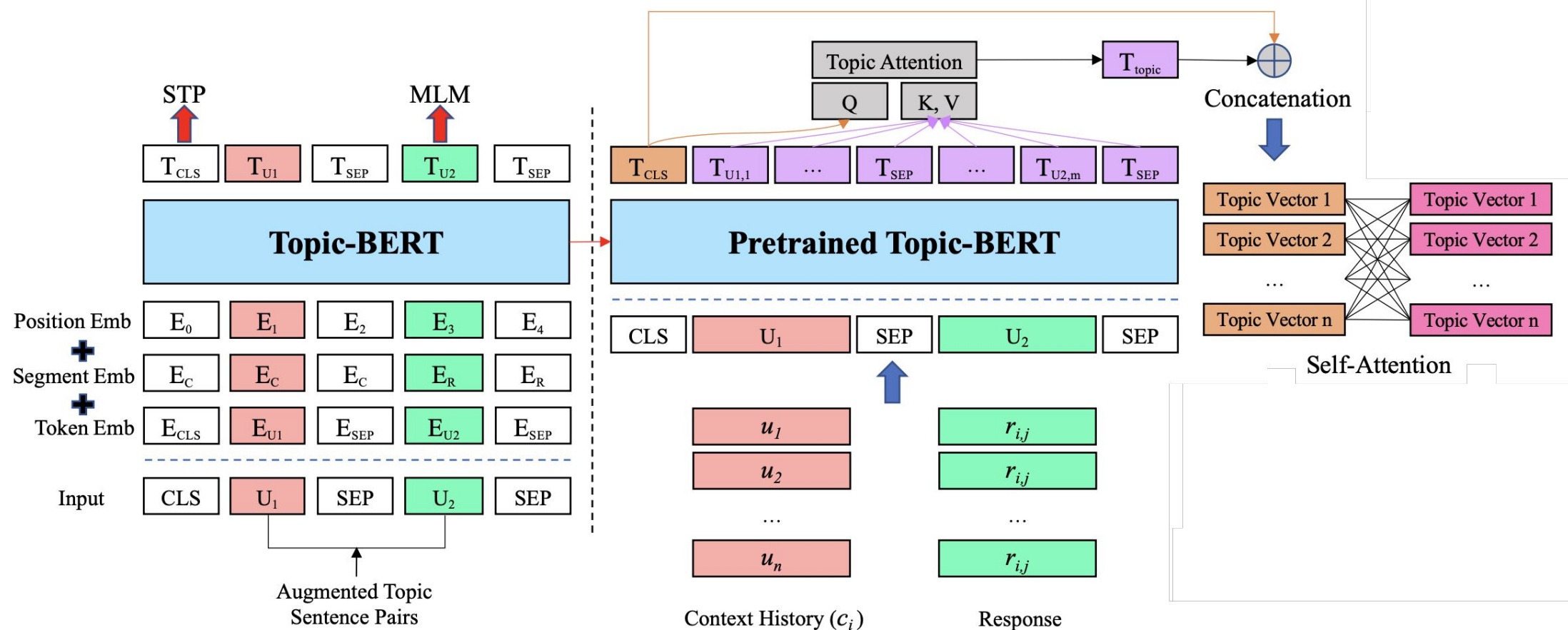
Topic attention to enhance topic information via [CLS] as query attending to the remaining tokens

Frame response selection as **dynamic topic tracking**



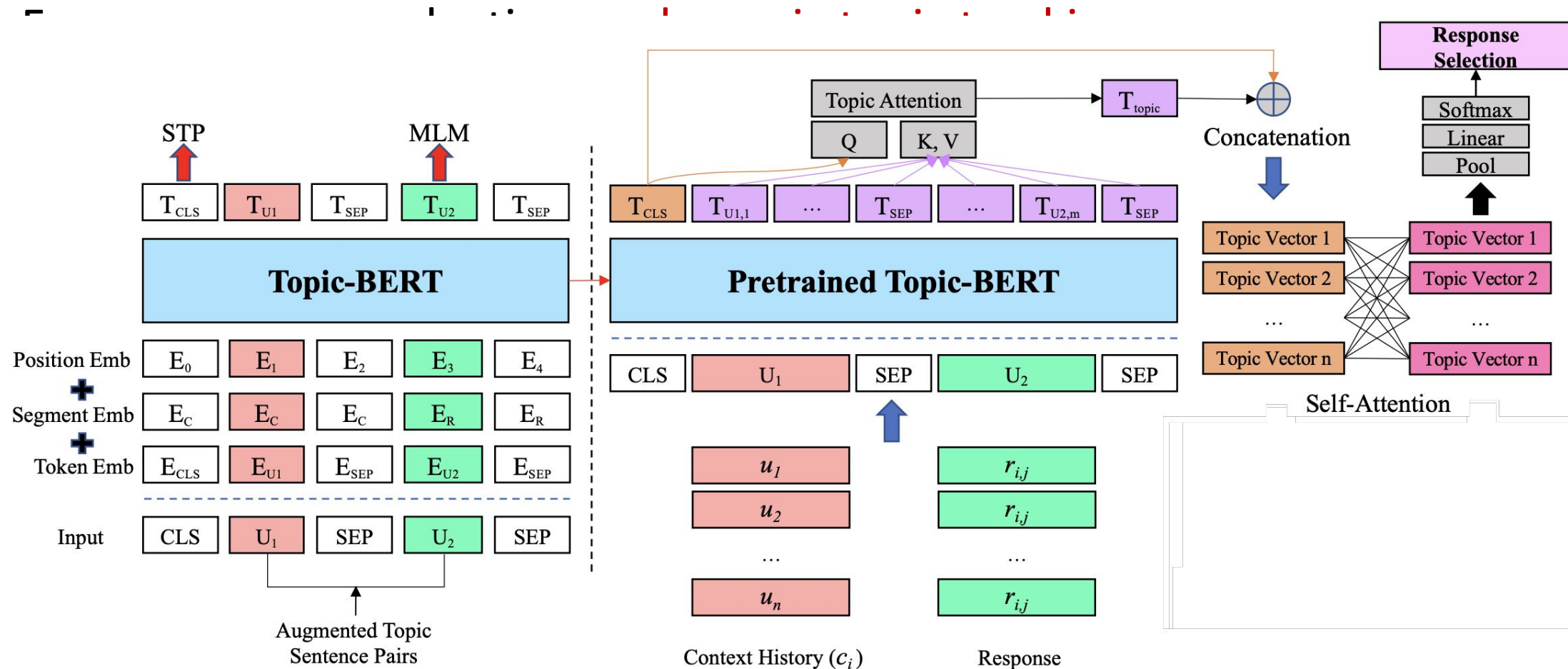
Topic-BERT

Operations for a response candidate with **each** conversation utterance

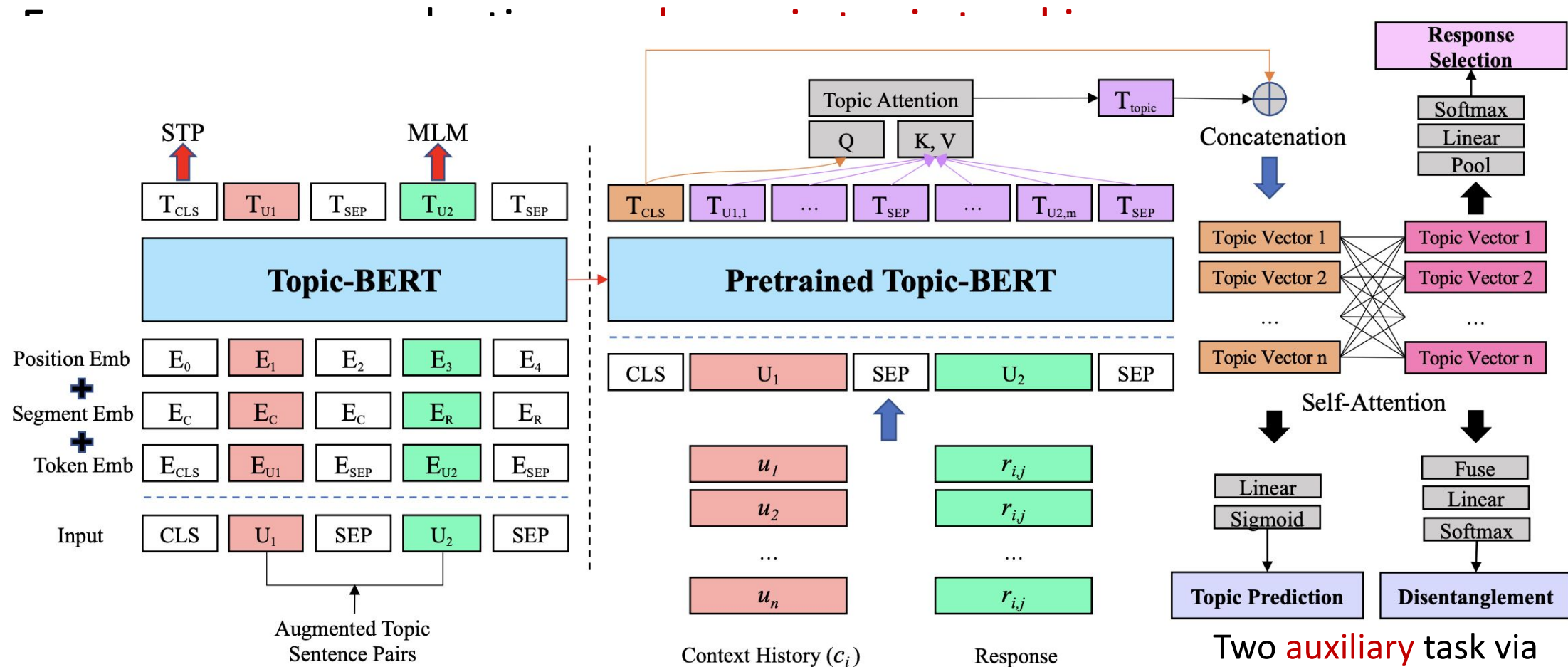


Topic-BERT

Response selection
as the **main task**



Topic-BERT

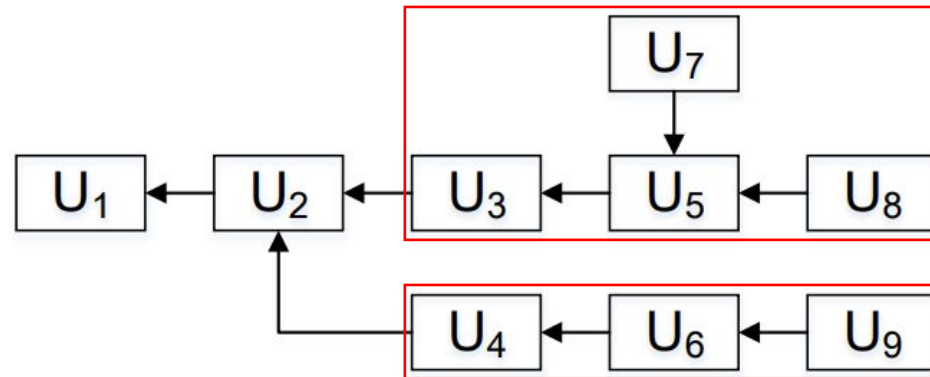


MPC-BERT: Utterance Semantics Modeling (1)

Shared Node Detection: a **full** MPC instance can be divided into **several sub-conversations**, e.g., two sub-conversations {U3, U5, U7, U8} and {U4, U6, U9} share the same **parent node U2**

MPC-BERT: Utterance Semantics Modeling (1)

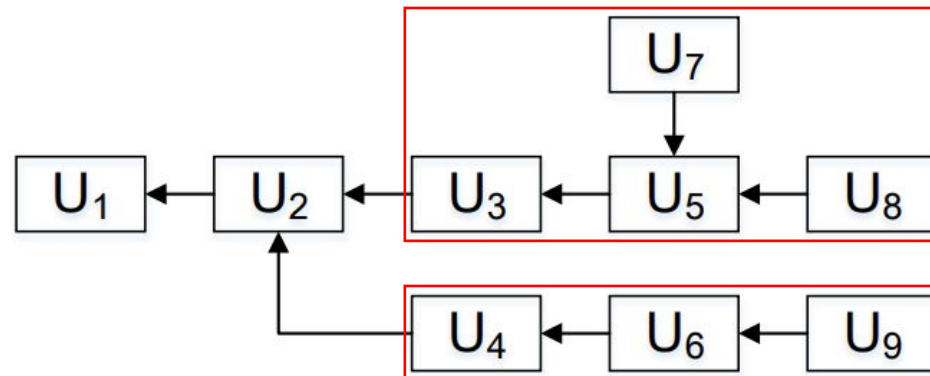
- **Shared Node Detection**: a full MPC instance can be divided into several sub-conversations, e.g., two sub-conversations {U3, U5, U7, U8} and {U4, U6, U9} share the same parent node U2



Interlocutor structure modeling has been covered in Section 3!

MPC-BERT: Utterance Semantics Modeling (1)

- **Shared Node Detection**: a **full** MPC instance can be divided into **several sub-conversations**, e.g., two sub-conversations $\{U_3, U_5, U_7, U_8\}$ and $\{U_4, U_6, U_9\}$ share the same **parent node** U_2



- **Assumption**: the representations of sub-conversations under **the same parent node** tend to be **similar**

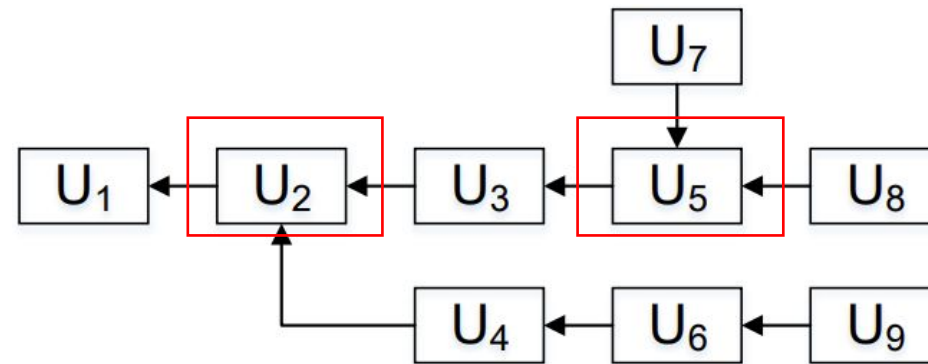
Interlocutor structure modeling has been covered in Section 3!

MPC-BERT: Utterance Semantics Modeling (2)

Masked Shared Utterance Restoration: a **shared** utterance is **semantically relevant to more utterances** in the context than non-shared ones, e.g., U2 and U5

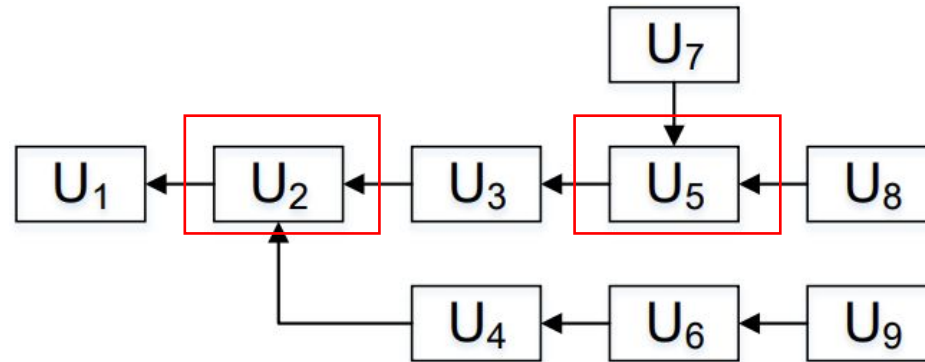
MPC-BERT: Utterance Semantics Modeling (2)

- **Masked Shared Utterance Restoration**: a **shared** utterance is **semantically relevant to more utterances** in the context than non-shared ones, e.g., U2 and U5



MPC-BERT: Utterance Semantics Modeling (2)

- **Masked Shared Utterance Restoration**: a **shared** utterance is **semantically relevant to more utterances** in the context than non-shared ones, e.g., U2 and U5



- **Assumption**: **mask** a sampled shared utterance and enforce model to **restore the masked shared utterance** given the rest conversation can enhance the conversation understanding

Results

MPC-BERT outperforms SA-BERT by margins of **3.82%, 2.71%, 2.55% and 3.22%** $R_{10}@1$

GIFT improve **BERT** by margins of **2.48%, 2.12%, 2.71% and 2.34%** $R_{10}@1$

	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$
DRNN (Ouchi and Tsuboi, 2016)	-	-	76.07	33.62	78.16	36.14	78.64	36.93
SIRNN (Zhang et al., 2018)	-	-	78.14	36.45	80.34	39.20	80.91	40.83
BERT (Devlin et al., 2019)	92.48	73.42	85.52	53.95	86.93	57.41	87.19	58.92
SA-BERT (Gu et al., 2020)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT (Gu et al., 2021)	94.90	78.98	87.63	57.95	89.14	61.82	89.70	63.64
BERT w/ GIFT	93.22 [†]	75.90 [†]	86.59 [†]	56.07 [†]	88.02 [†]	60.12 [†]	88.57 [†]	61.26 [†]
SA-BERT w/ GIFT	94.26 [†]	78.20 [†]	88.07[†]	59.40[†]	89.91[†]	64.45[†]	90.45 [†]	65.77 [†]
MPC-BERT w/ GIFT	95.04	80.74[†]	87.97	58.83 [†]	89.77 [†]	63.97 [†]	90.62[†]	66.08[†]

GIFT is also evaluated and shows effectiveness on response selection ⁷¹

Results

MPC-BERT outperforms SA-BERT by margins of 3.82%, 2.71%, 2.55% and 3.22% $R_{10}@1$

GIFT improve BERT by margins of 2.48%, 2.12%, 2.71% and 2.34% $R_{10}@1$

	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$
DRNN (Ouchi and Tsuboi, 2016)	-	-	76.07	33.62	78.16	36.14	78.64	36.93
SIRNN (Zhang et al., 2018)	-	-	78.14	36.45	80.34	39.20	80.91	40.83
BERT (Devlin et al., 2019)	92.48	73.42	85.52	53.95	86.93	57.41	87.19	58.92
SA-BERT (Gu et al., 2020)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT (Gu et al., 2021)	94.90	78.98	87.63	57.95	89.14	61.82	89.70	63.64
BERT w/ GIFT	93.22 [†]	75.90 [†]	86.59 [†]	56.07 [†]	88.02 [†]	60.12 [†]	88.57 [†]	61.26 [†]
SA-BERT w/ GIFT	94.26 [†]	78.20 [†]	88.07[†]	59.40[†]	89.91[†]	64.45[†]	90.45 [†]	65.77 [†]
MPC-BERT w/ GIFT	95.04	80.74[†]	87.97	58.83 [†]	89.77 [†]	63.97 [†]	90.62[†]	66.08[†]

GIFT is also evaluated and shows effectiveness on response selection ⁷¹

Results

MPC-BERT outperforms SA-BERT by margins of 3.82%, 2.71%, 2.55% and 3.22% $R_{10}@1$

GIFT improve BERT by margins of 2.48%, 2.12%, 2.71% and 2.34% $R_{10}@1$

	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$
DRNN (Ouchi and Tsuboi, 2016)	-	-	76.07	33.62	78.16	36.14	78.64	36.93
SIRNN (Zhang et al., 2018)	-	-	78.14	36.45	80.34	39.20	80.91	40.83
BERT (Devlin et al., 2019)	92.48	73.42	85.52	53.95	86.93	57.41	87.19	58.92
SA-BERT (Gu et al., 2020)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT (Gu et al., 2021)	94.90	78.98	87.63	57.95	89.14	61.82	89.70	63.64
BERT w/ GIFT	93.22 [†]	75.90 [†]	86.59 [†]	56.07 [†]	88.02 [†]	60.12 [†]	88.57 [†]	61.26 [†]
SA-BERT w/ GIFT	94.26 [†]	78.20 [†]	88.07[†]	59.40[†]	89.91[†]	64.45[†]	90.45 [†]	65.77 [†]
MPC-BERT w/ GIFT	95.04	80.74[†]	87.97	58.83 [†]	89.77 [†]	63.97 [†]	90.62[†]	66.08[†]

GIFT is also evaluated and shows effectiveness on response selection ⁷¹

Results

MPC-BERT outperforms SA-BERT by margins of 3.82%, 2.71%, 2.55% and 3.22% $R_{10}@1$

GIFT improve BERT by margins of 2.48%, 2.12%, 2.71% and 2.34% $R_{10}@1$

improve SA-BERT by margins of 3.04%, 4.16%, 5.18% and 5.35% $R_{10}@1$

	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$
DRNN (Ouchi and Tsuboi, 2016)	-	-	76.07	33.62	78.16	36.14	78.64	36.93
SIRNN (Zhang et al., 2018)	-	-	78.14	36.45	80.34	39.20	80.91	40.83
BERT (Devlin et al., 2019)	92.48	73.42	85.52	53.95	86.93	57.41	87.19	58.92
SA-BERT (Gu et al., 2020)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT (Gu et al., 2021)	94.90	78.98	87.63	57.95	89.14	61.82	89.70	63.64
BERT w/ GIFT	93.22 [†]	75.90 [†]	86.59 [†]	56.07 [†]	88.02 [†]	60.12 [†]	88.57 [†]	61.26 [†]
SA-BERT w/ GIFT	94.26 [†]	78.20 [†]	88.07[†]	59.40[†]	89.91[†]	64.45[†]	90.45 [†]	65.77 [†]
MPC-BERT w/ GIFT	95.04	80.74[†]	87.97	58.83 [†]	89.77 [†]	63.97 [†]	90.62[†]	66.08[†]

GIFT is also evaluated and shows effectiveness on response selection ⁷¹

Results

MPC-BERT outperforms SA-BERT by margins of 3.82%, 2.71%, 2.55% and 3.22% $R_{10}@1$

GIFT improve BERT by margins of 2.48%, 2.12%, 2.71% and 2.34% $R_{10}@1$

improve SA-BERT by margins of 3.04%, 4.16%, 5.18% and 5.35% $R_{10}@1$

improve MPC-BERT by margins of 1.76%, 0.88%, 2.15% and 2.44% $R_{10}@1$

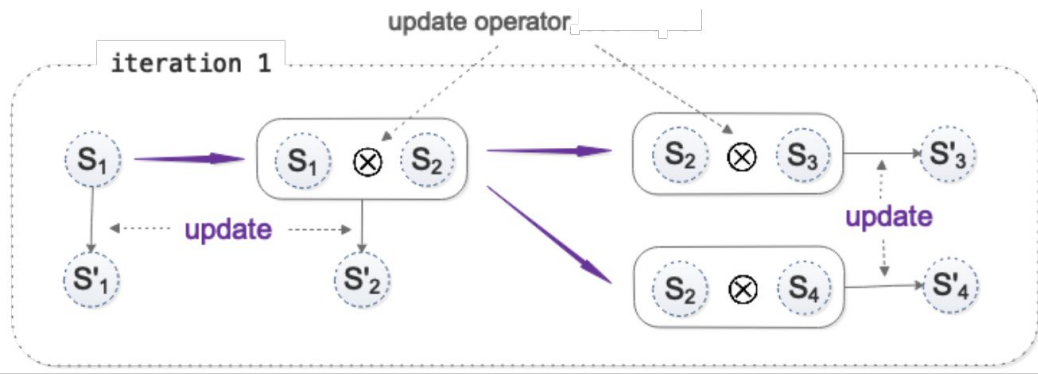
	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$	$R_2@1$	$R_{10}@1$
DRNN (Ouchi and Tsuboi, 2016)	-	-	76.07	33.62	78.16	36.14	78.64	36.93
SIRNN (Zhang et al., 2018)	-	-	78.14	36.45	80.34	39.20	80.91	40.83
BERT (Devlin et al., 2019)	92.48	73.42	85.52	53.95	86.93	57.41	87.19	58.92
SA-BERT (Gu et al., 2020)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT (Gu et al., 2021)	94.90	78.98	87.63	57.95	89.14	61.82	89.70	63.64
BERT w/ GIFT	93.22 [†]	75.90 [†]	86.59 [†]	56.07 [†]	88.02 [†]	60.12 [†]	88.57 [†]	61.26 [†]
SA-BERT w/ GIFT	94.26 [†]	78.20 [†]	88.07[†]	59.40[†]	89.91[†]	64.45[†]	90.45 [†]	65.77 [†]
MPC-BERT w/ GIFT	95.04	80.74[†]	87.97	58.83 [†]	89.77 [†]	63.97 [†]	90.62[†]	66.08[†]

GIFT is also evaluated and shows effectiveness on response selection ⁷¹

Representative Tasks

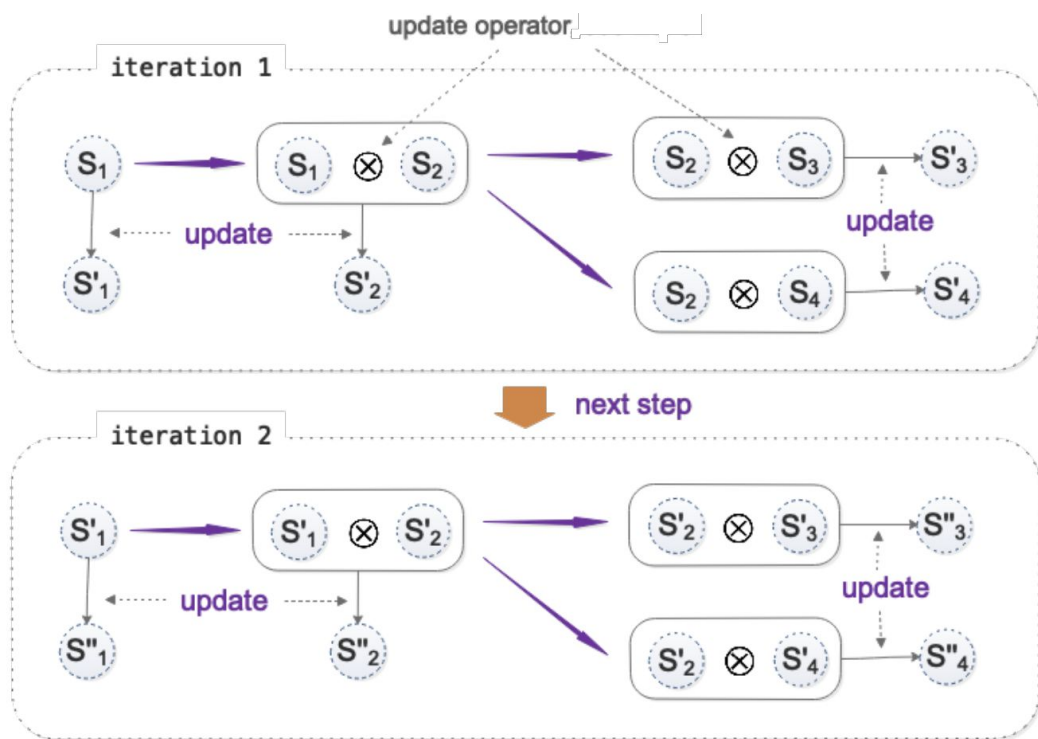
- Response selection aims at selecting the best-matched response from a set of candidates, given the context of a multi-turn conversation (retrieval-based)
- **Response generation** synthesize a response with a natural language **generative model** by maximizing its generation probability given the previous conversation history (**generation**-based)

GSN



Utterance-level graph-based encoder which encodes utterances based on the graph topology rather than the appearance sequence

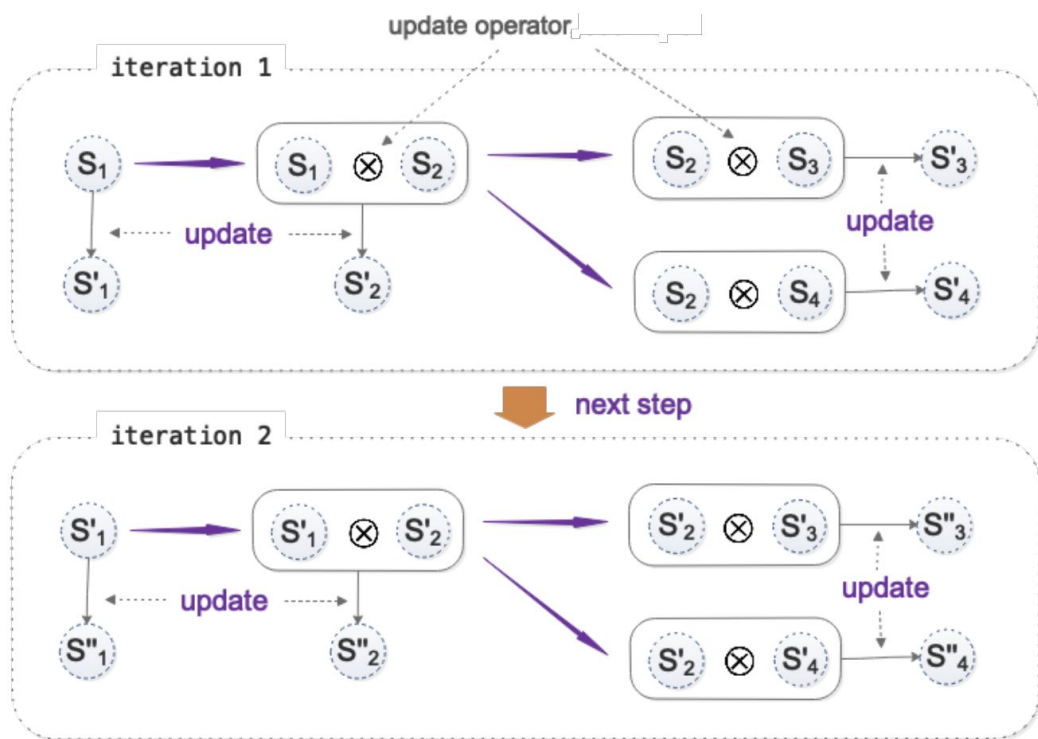
GSN



Utterance-level **graph**-based encoder which encodes utterances based on the **graph topology** rather than the appearance sequence

Each **utterance** (a **node** in the graph) accepts information from all its **connected utterances** (nodes) in each iteration

GSN

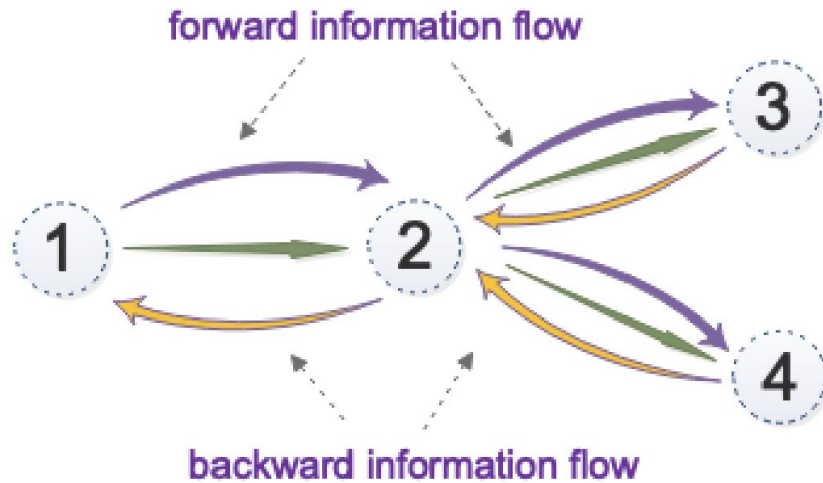


Utterance-level **graph**-based encoder which encodes utterances based on the **graph topology** rather than the appearance sequence

Each **utterance** (a **node** in the graph) accepts information from all its **connected utterances** (nodes) in each iteration

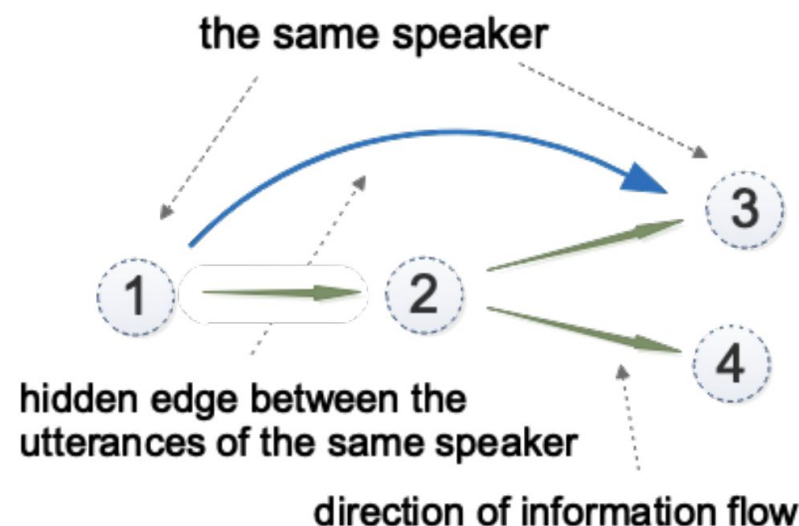
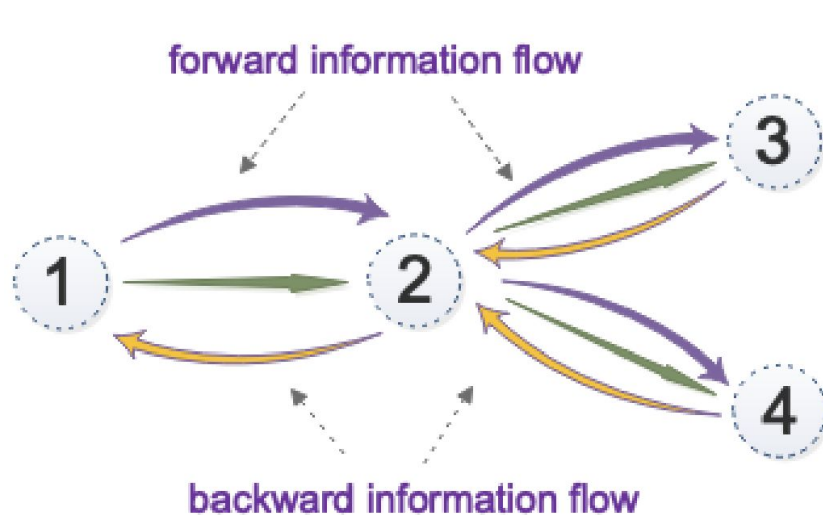
Homogeneous graph
composed of only
utterances!

Bi-directional & Speaker Information Flow



Bi-directional: to allow information to flow **thoroughly** via **backward** and **forward** propagation

Bi-directional & Speaker Information Flow



Bi-directional: to allow information to flow **thoroughly** via **backward** and **forward** propagation

Speaker: to reflect **speaker change** via creating an edge for every **utterance pair** from the **same speaker**

Is a homogeneous graph expressive enough to represent an MPC?



Is a homogeneous graph expressive enough to represent an MPC?



Q1 Are there **other sources of information** in addition to utterances that should be embraced in a unified graph?

Is a homogeneous graph expressive enough to represent an MPC?



Q1 Are there **other sources of information** in addition to utterances that should be embraced in a unified graph?

Q2 Is it necessary to distinguish the **fine-grained** and **complicated interactions** between utterance and interlocutor graph nodes?

HeterMPC: Graph Construction

- M utterances and I interlocutors $\square$ a heterogeneous graph $G(V, E)$

HeterMPC: Graph Construction

- M utterances and I interlocutors $\square$ a **heterogeneous** graph $G(V, E)$
- V : a set of $M + I$ nodes, each denoting an **utterance** or an **interlocutor**

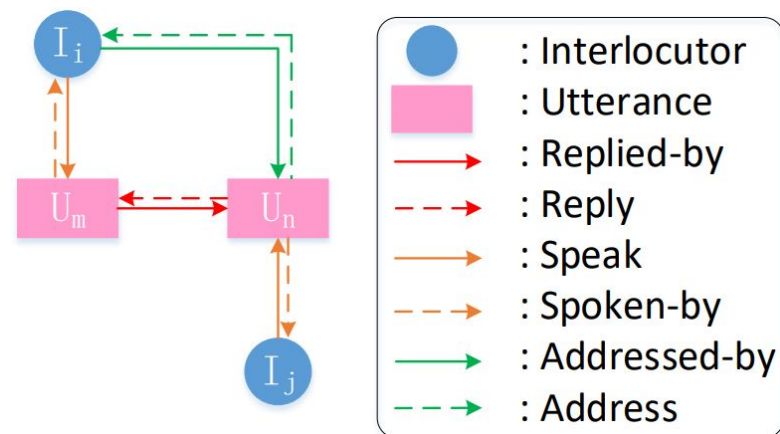
HeterMPC: Graph Construction

- M utterances and I interlocutors $\square$ a **heterogeneous** graph $G(V, E)$
- V : a set of $M + I$ nodes, each denoting an **utterance** or an **interlocutor**
- $E = \{e_{p,q}\}_{p,q=1}^{M+I}$: a set of **directed edges**, each edge $e_{p,q}$ describing the connection from node p to node q

HeterMPC: Graph Construction

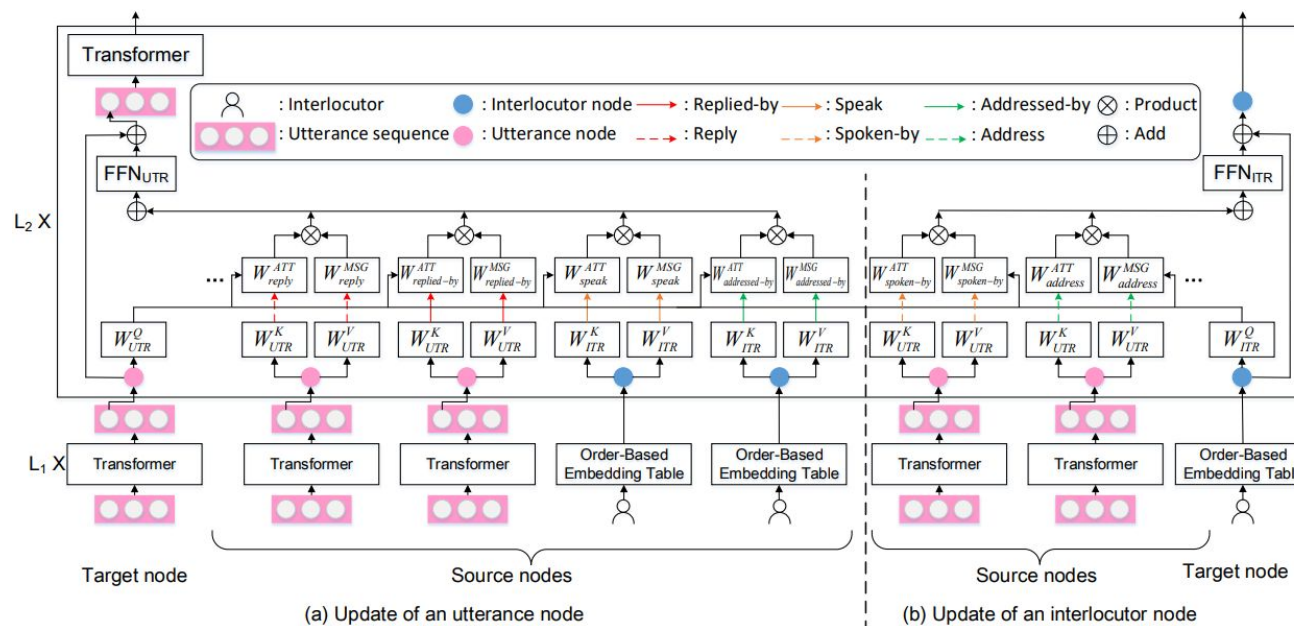
- M utterances and I interlocutors $\square$ a **heterogeneous** graph $G(V, E)$
- V : a set of $M + I$ nodes, each denoting an **utterance** or an **interlocutor**
- $E = \{e_{p,q}\}_{p,q=1}^{M+I}$: a set of **directed edges**, each edge $e_{p,q}$ describing the connection from node p to node q

- Six types of meta relations: {**reply**, **replied-by**, **speak**, **spoken-by**, **address**, **addressed-by**} to describe directed edges between two nodes



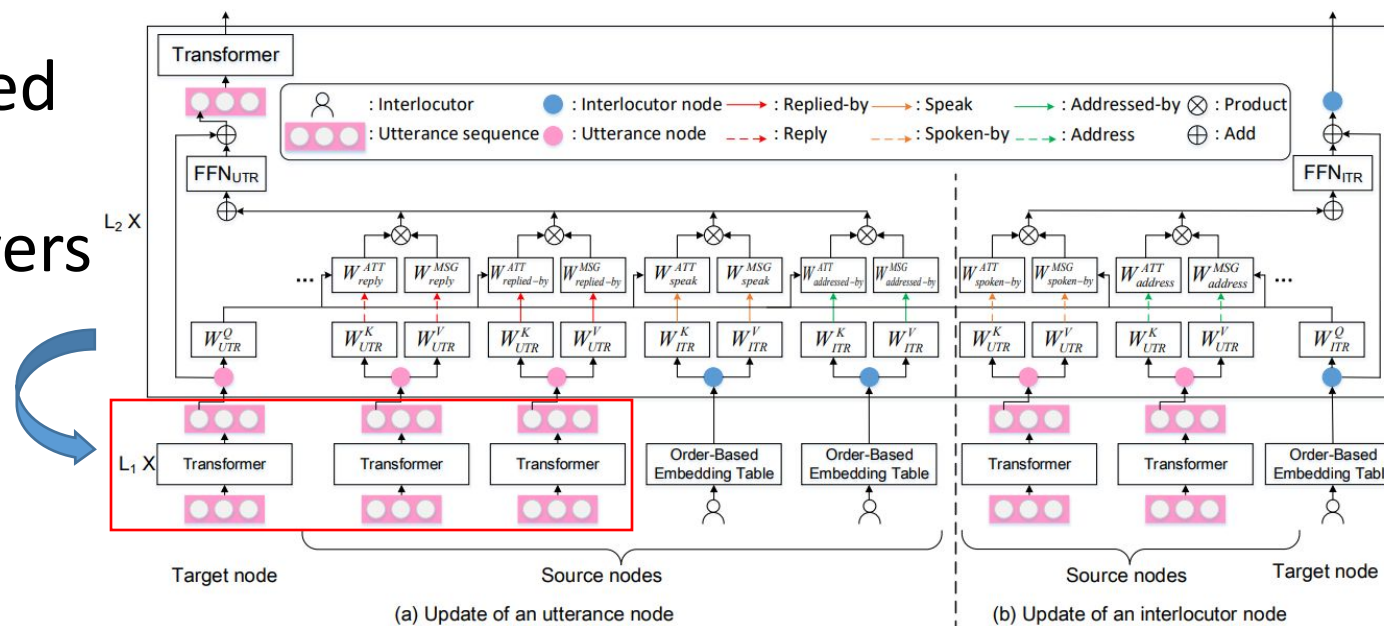
HeterMPC: Node Initialization

Each **utterance** is encoded individually by stacked **Transformer encoder** layers



HeterMPC: Node Initialization

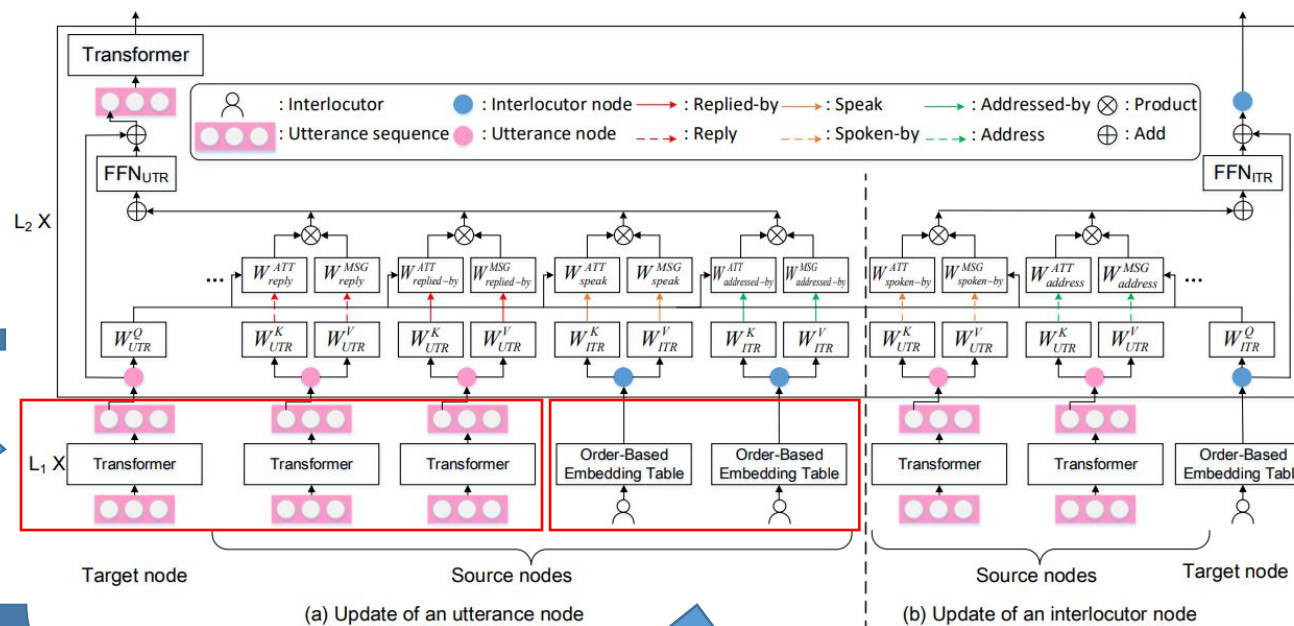
- Each **utterance** is encoded individually by stacked **Transformer encoder** layers



HeterMPC: Node Initialization

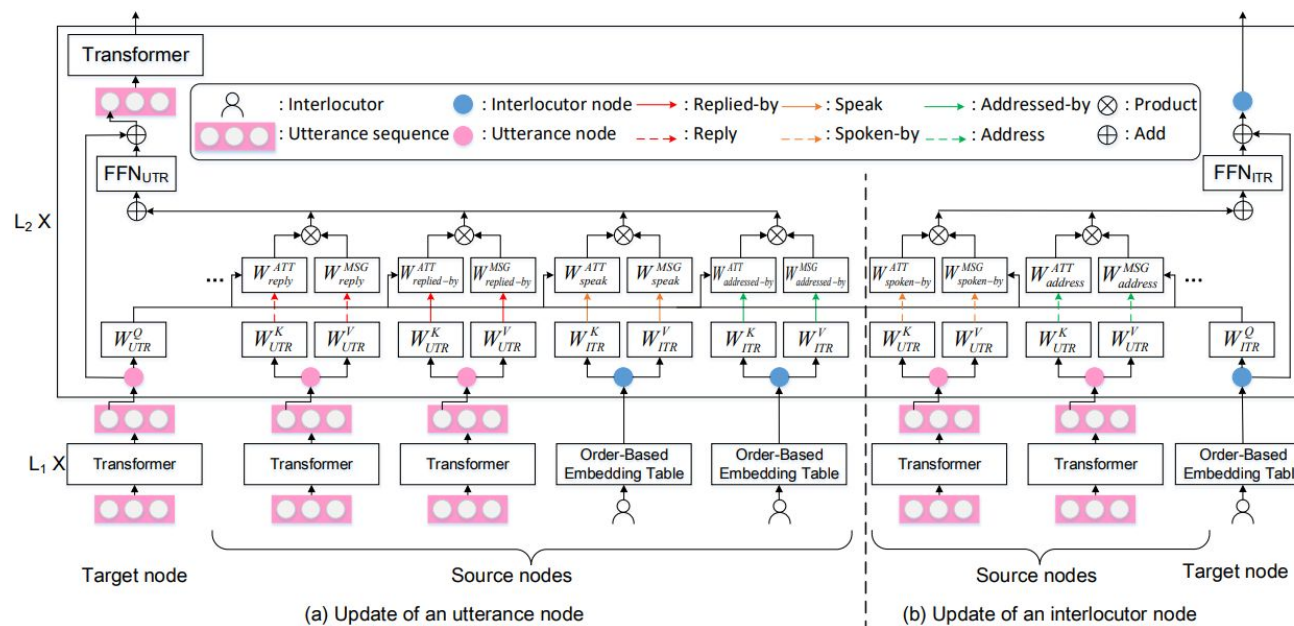
- Each **utterance** is encoded individually by stacked **Transformer encoder** layers

- Each **interlocutor** is directly represented by looking up a **position-based interlocutor embedding table**



HeterMPC: Node Updating

Introduce **parameters** to model **heterogeneity** via



HeterMPC: Node Updating

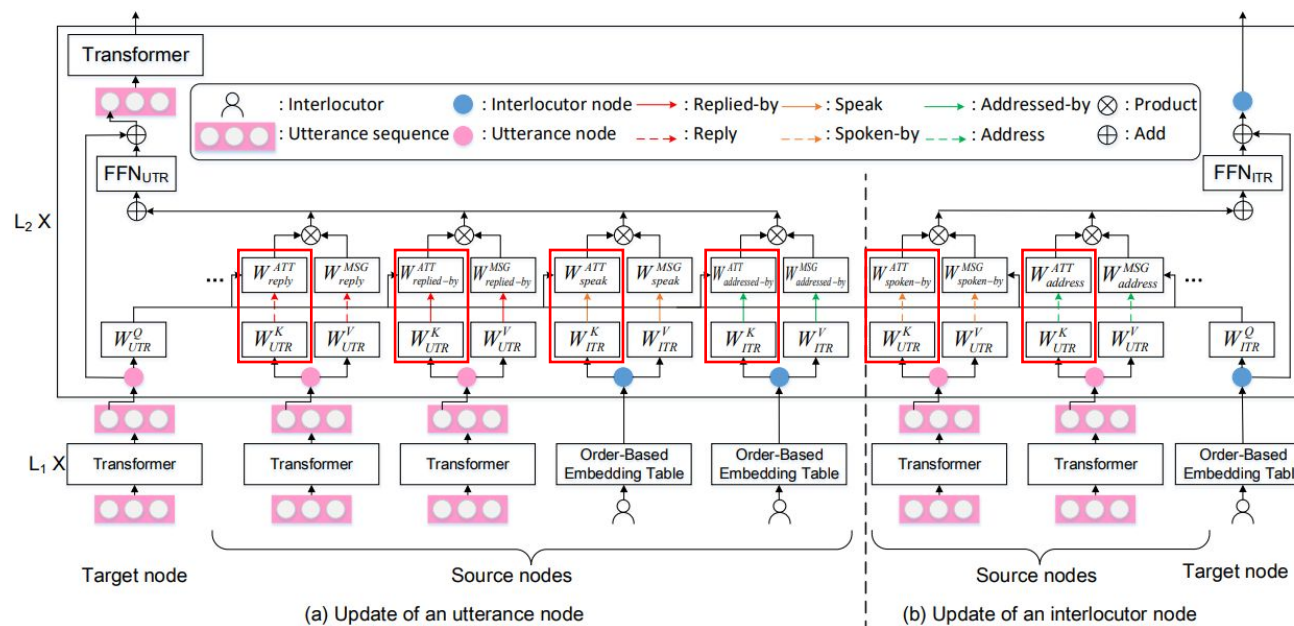
Introduce **parameters** to model **heterogeneity** via

- **attention weights**

$$\mathbf{k}^l(s) = \mathbf{h}_s^l \mathbf{W}_{\tau(s)}^K + \mathbf{b}_{\tau(s)}^K,$$

$$\mathbf{q}^l(t) = \mathbf{h}_t^l \mathbf{W}_{\tau(t)}^Q + \mathbf{b}_{\tau(t)}^Q,$$

$$\mathbf{w}^l(s, e, t) = \mathbf{k}^l(s) \mathbf{W}_{e_s, t}^{ATT} \mathbf{q}^l(t)^T \frac{\mu_{e_s, t}}{\sqrt{d}}.$$



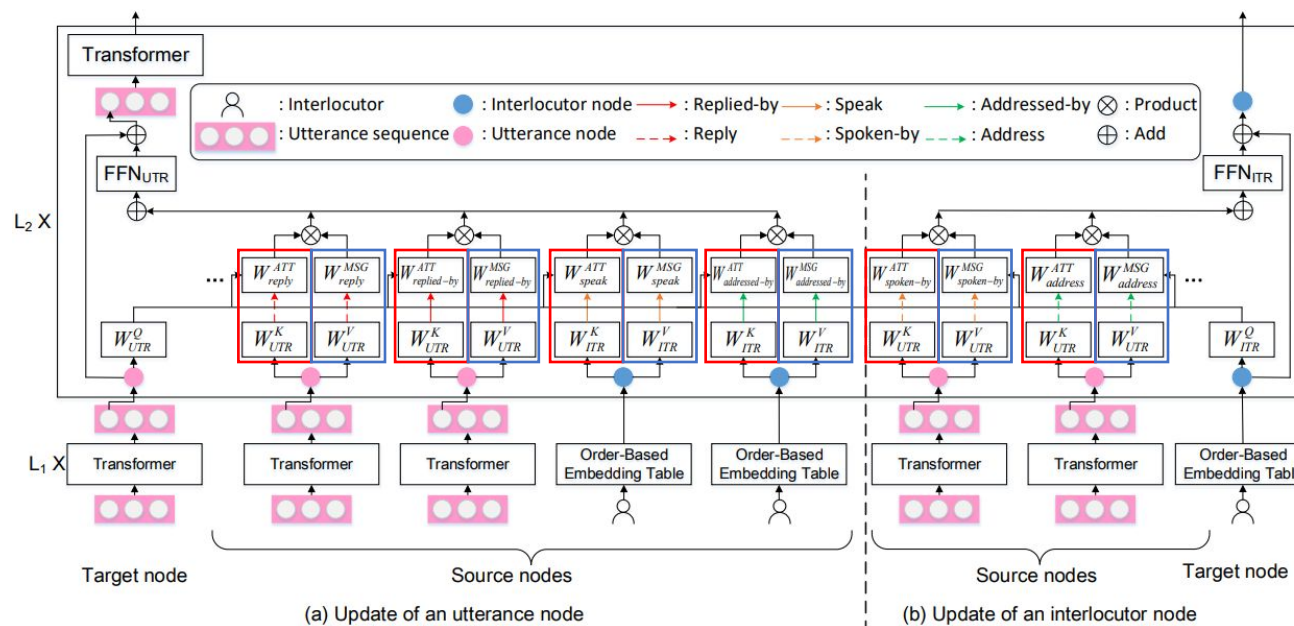
HeterMPC: Node Updating

Introduce **parameters** to model **heterogeneity** via

- **attention weights**
- **message passing**

$$\mathbf{v}^l(s) = \mathbf{h}_s^l \mathbf{W}_{\tau(s)}^V + \mathbf{b}_{\tau(s)}^V,$$

$$\bar{\mathbf{v}}^l(s) = \mathbf{v}^l(s) \mathbf{W}_{e_{s,t}}^{MSG},$$



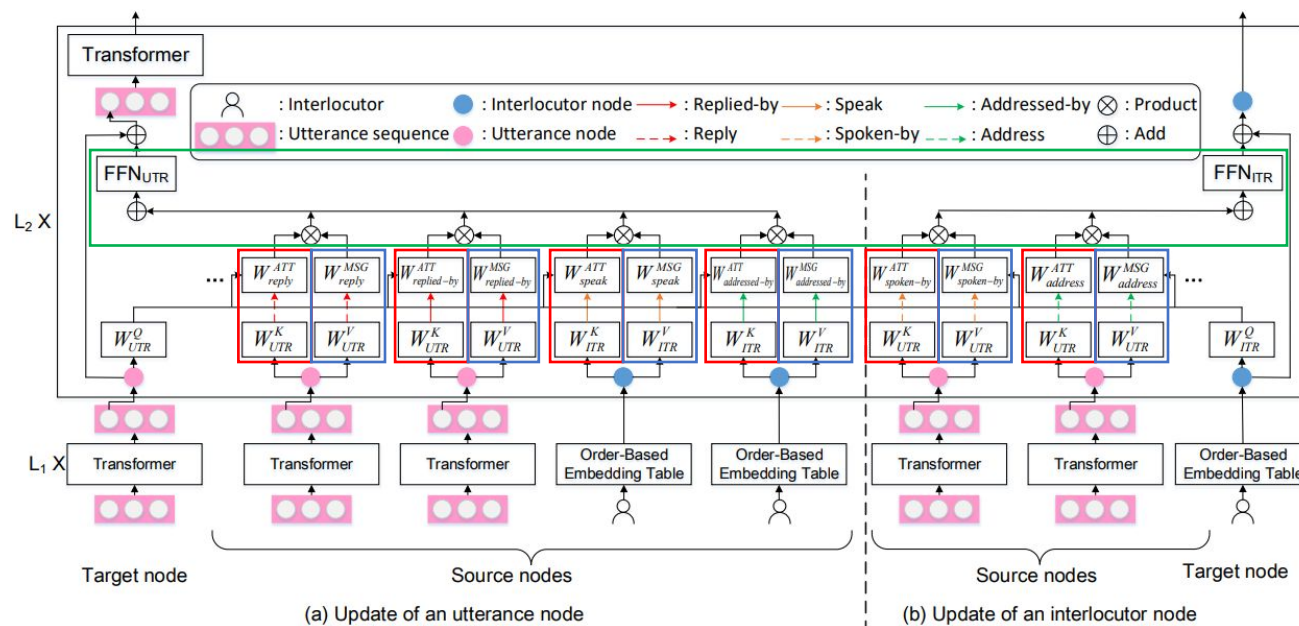
HeterMPC: Node Updating

Introduce **parameters** to model **heterogeneity** via

- **attention weights**
- **message passing**
- **information aggregation**

$$\bar{h}_t^l = \sum \text{softmax}(w^l(s, e, t)) \bar{v}^l(s),$$

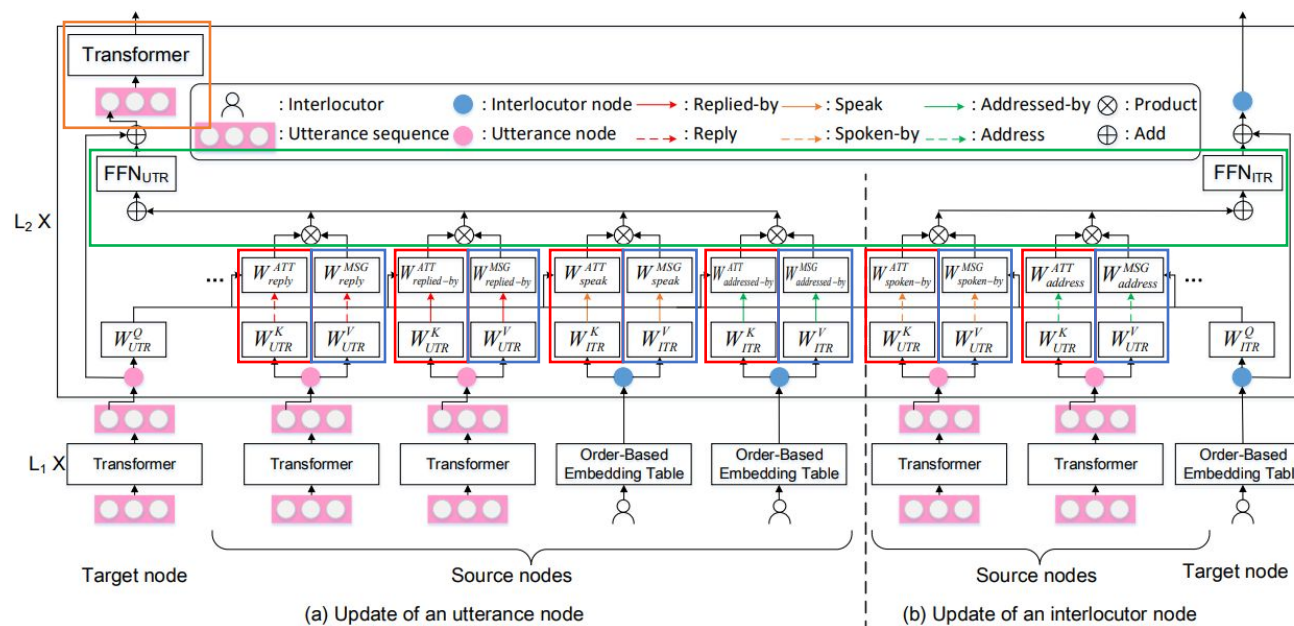
$$h_t^{l+1} = \text{FFN}_{\tau(t)}(\bar{h}_t^l) + h_t^l,$$



HeterMPC: Node Updating

Introduce **parameters** to model **heterogeneity** via

- **attention weights**
- **message passing**
- **information aggregation**



- Specifically, the context information in an **utterance node** is shared with **other tokens in this utterance** through another layer of intra-utterance Transformer encoding

Graph-based methods heavily rely on the necessary addressee labels to construct a consecutively connected conversation graph

Graph-based methods heavily rely on the necessary addressee labels to construct a consecutively connected conversation graph

Speaker	Utterance	Addressee
User 1	"Good point, tmux is the thing I miss."	–
User 1	"Cool thanks for ur help." @User 4	User 4
User 2	"Ahha, you r using something like cpanel."	–
User 3	"Yeah 1.4.0 exactly." @User 2	User 2
User 4	"my pleasure :)"	–

Scarcity of addressee labels: addressees of 55% of the utterances in the Ubuntu IRC dataset (Ouchi and Tsuboi, 2016) are not specified

Graph-based methods heavily rely on the necessary addressee labels to construct a consecutively connected conversation graph

Speaker	Utterance	Addressee
User 1	"Good point, tmux is the thing I miss."	–
User 1	"Cool thanks for ur help." @User 4	User 4
User 2	"Ahha, you r using something like cpanel."	–
User 3	"Yeah 1.4.0 exactly." @User 2	User 2
User 4	"my pleasure :)"	–

Scarcity of addressee labels: addressees of 55% of the utterances in the Ubuntu IRC dataset (Ouchi and Tsuboi, 2016) are not specified

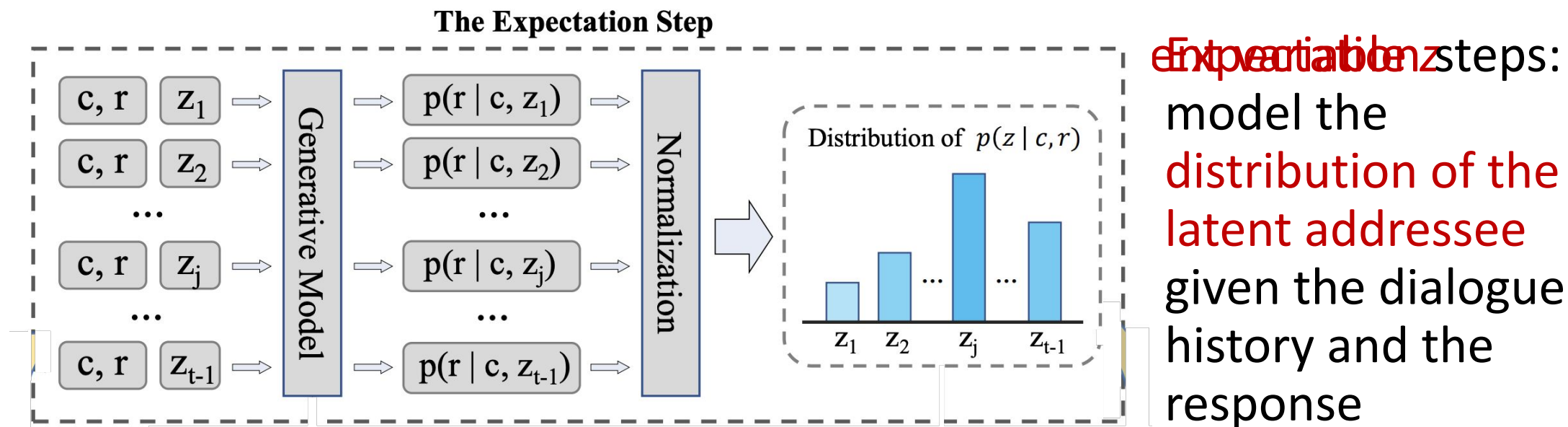


Result in only a few separate conversation fragments

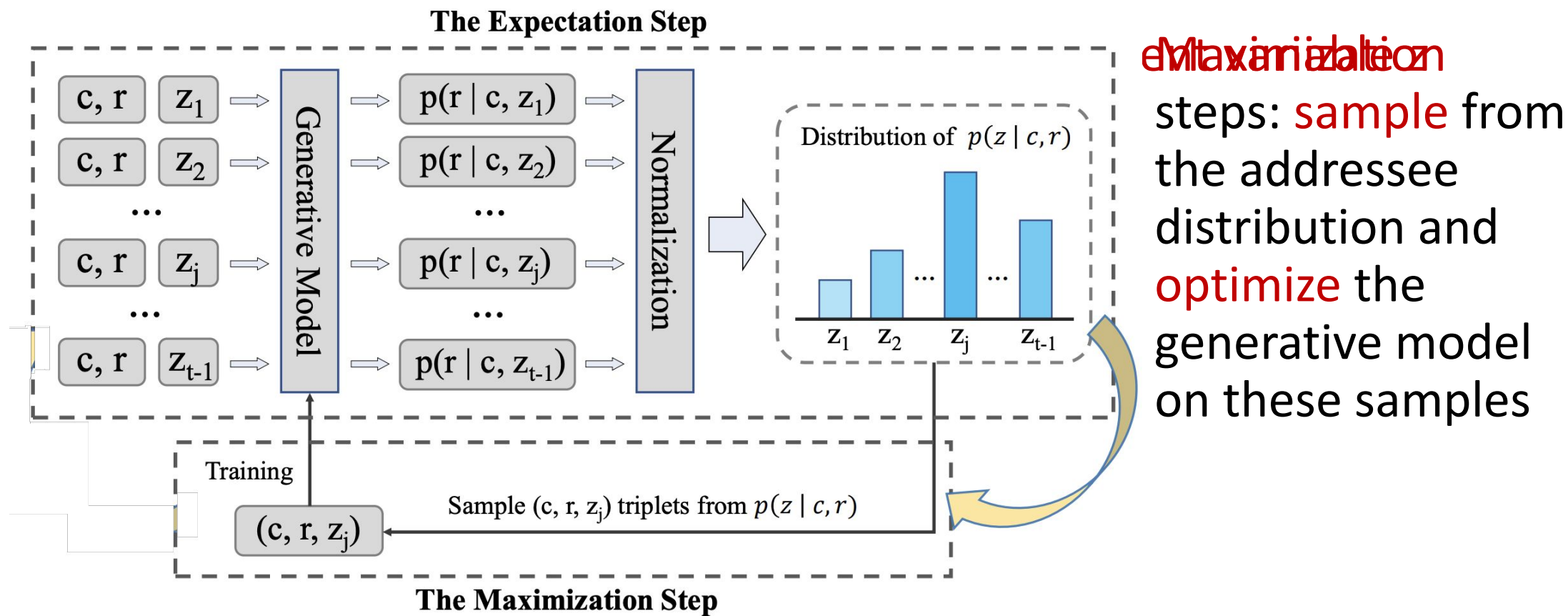
EM Pre-training

Treat the addressee of an utterance as a discrete **latent variable z**

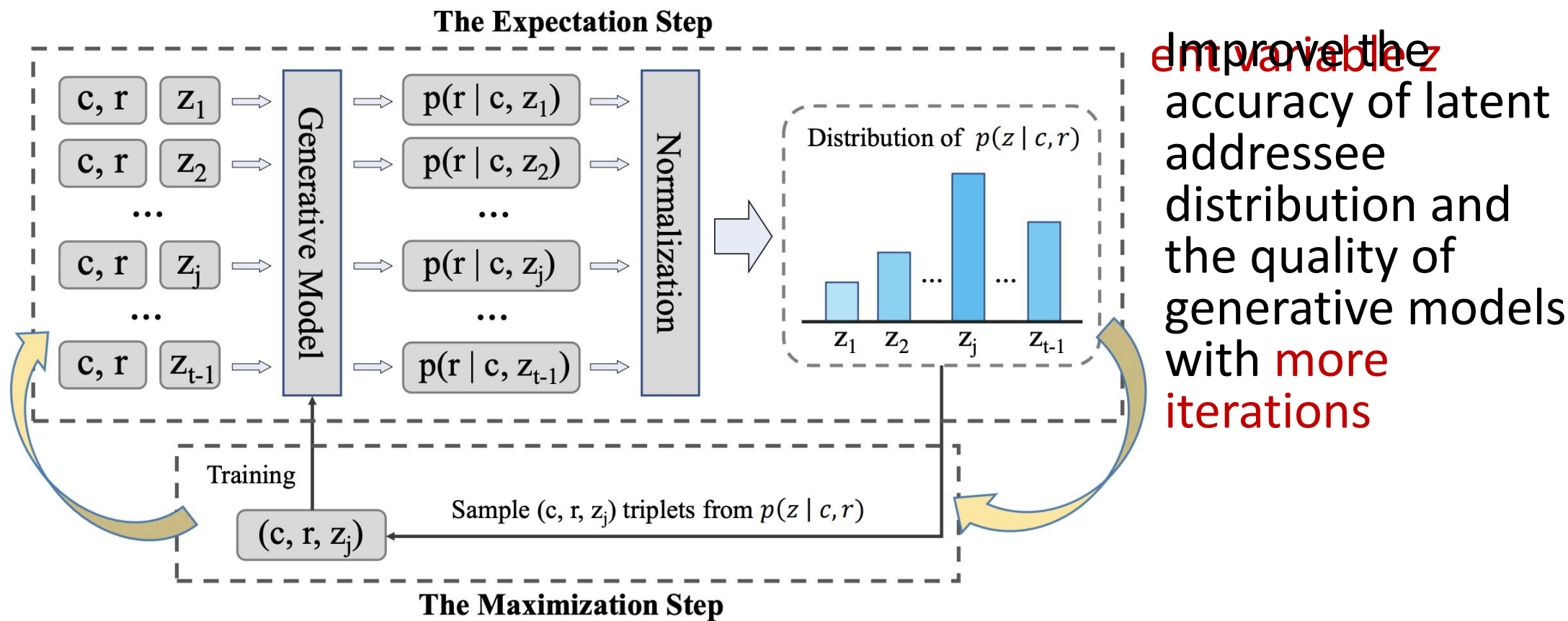
EM Pre-training



EM Pre-training



EM Pre-training



Results

Graph-based outperforms non-graph-based

Heterogeneous-graph-based outperforms homogeneous-graph-based

Addressee-filled outperforms addressee-missed

Models	Metrics					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE _L
RNN Seq2Seq (Sutskever et al., 2014)	7.71	2.46	1.12	0.64	3.33	8.68
Transformer (Vaswani et al., 2017)	7.89	2.75	1.34	0.74	3.81	9.20
GSN (Hu et al., 2019)	10.23	3.57	1.70	0.97	4.10	9.91
GPT-2 (Radford et al., 2019)	10.37	3.60	1.66	0.93	4.01	9.53
BART (Lewis et al., 2020)	11.25	4.02	1.78	0.95	4.46	9.90
HeterMPC (Gu et al., 2022)	12.26	4.80	2.42	1.49	4.94	11.20
EM Pretraining (Li and Zhao, 2023)	12.31	5.39	3.34	2.45	5.52	11.71

Results

Graph-based outperforms non-graph-based

Heterogeneous-graph-based outperforms homogeneous-graph-based

Addressee-filled outperforms addressee-missed

Models \ Metrics	Metrics					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE _L
RNN Seq2Seq (Sutskever et al., 2014)	7.71	2.46	1.12	0.64	3.33	8.68
Transformer (Vaswani et al., 2017)	7.89	2.75	1.34	0.74	3.81	9.20
GSN (Hu et al., 2019)	10.23	3.57	1.70	0.97	4.10	9.91
GPT-2 (Radford et al., 2019)	10.37	3.60	1.66	0.93	4.01	9.53
BART (Lewis et al., 2020)	11.25	4.02	1.78	0.95	4.46	9.90
HeterMPC (Gu et al., 2022)	12.26	4.80	2.42	1.49	4.94	11.20
EM Pretraining (Li and Zhao, 2023)	12.31	5.39	3.34	2.45	5.52	11.71

Results

Graph-based outperforms non-graph-based

Heterogeneous-graph-based outperforms homogeneous-graph-based

Addressee-filled outperforms addressee-missed

Models	Metrics					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE _L
RNN Seq2Seq (Sutskever et al., 2014)	7.71	2.46	1.12	0.64	3.33	8.68
Transformer (Vaswani et al., 2017)	7.89	2.75	1.34	0.74	3.81	9.20
GSN (Hu et al., 2019)	10.23	3.57	1.70	0.97	4.10	9.91
GPT-2 (Radford et al., 2019)	10.37	3.60	1.66	0.93	4.01	9.53
BART (Lewis et al., 2020)	11.25	4.02	1.78	0.95	4.46	9.90
HeterMPC (Gu et al., 2022)	12.26	4.80	2.42	1.49	4.94	11.20
EM Pretraining (Li and Zhao, 2023)	12.31	5.39	3.34	2.45	5.52	11.71

Results

Graph-based outperforms non-graph-based

Heterogeneous-graph-based outperforms homogeneous-graph-based

Addressee-filled outperforms addressee-missed

Models	Metrics					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE _L
RNN Seq2Seq (Sutskever et al., 2014)	7.71	2.46	1.12	0.64	3.33	8.68
Transformer (Vaswani et al., 2017)	7.89	2.75	1.34	0.74	3.81	9.20
GSN (Hu et al., 2019)	10.23	3.57	1.70	0.97	4.10	9.91
GPT-2 (Radford et al., 2019)	10.37	3.60	1.66	0.93	4.01	9.53
BART (Lewis et al., 2020)	11.25	4.02	1.78	0.95	4.46	9.90
HeterMPC (Gu et al., 2022)	12.26	4.80	2.42	1.49	4.94	11.20
EM Pretraining (Li and Zhao, 2023)	12.31	5.39	3.34	2.45	5.52	11.71

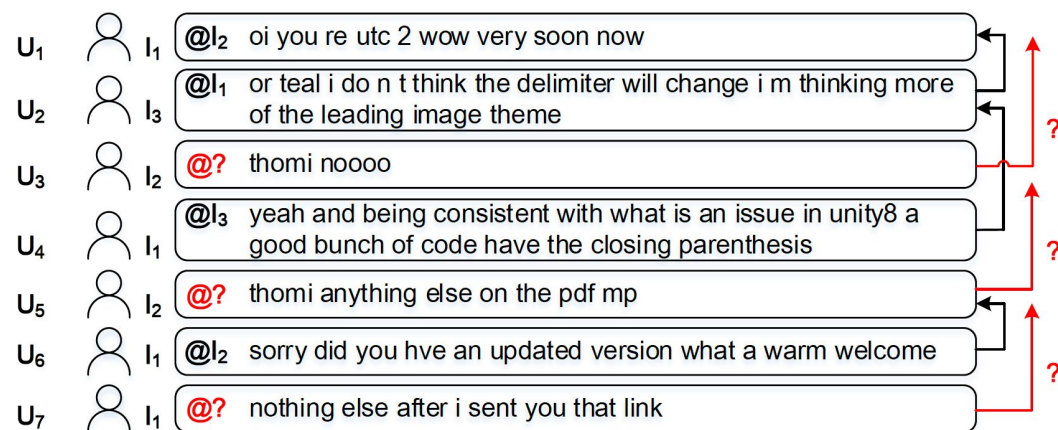
Section 6: Challenges & Opportunities

Tutorial Summary

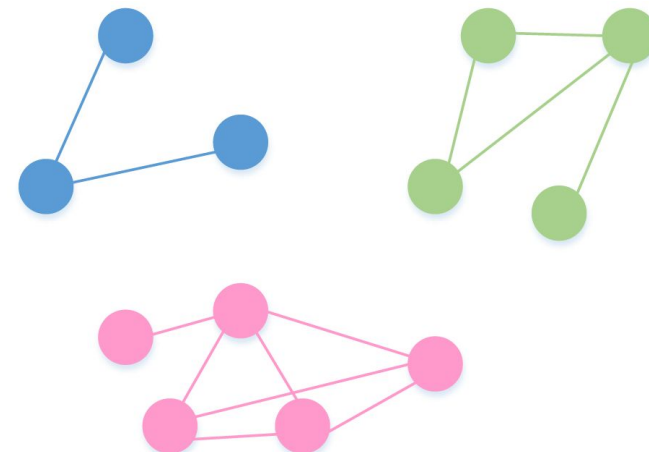
- **WHO** speaks
 - **Turn-taking** to determine if take the floor to speak or not
 - **Speaker identification** to identify the speaker of a specific utterance
- Address **WHOM**
 - ✓ **Explicit** addressee recognition to recognize the addressee of a specific utterance
 - ✓ **Implicit** dialogue disentanglement to disentangle a whole conversation from a data stream into several threads
- Say **WHAT**
 - ✓ **Retrieval**-based to rank a list of response candidates
 - ✓ **Generation**-based to synthesize a response via generative models

Challenge (1): Scarcity of Addressee Labels

Reduce the heavy dependency on the necessary addressee labels



(a) An MPC instance with a few addressee labels (@) missing

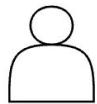


Addressees of **55%** of the utterances in the Ubuntu dataset are **not specified**

Fail to build a **consecutively connected** conversation graph, but only a few **separate conversation fragments** instead

Challenge (2): ChatMPC Solver

Is ChatGPT a Good Multi-Party Conversation Solver?



Human

Can you solve multi-party conversation tasks?

I do not have built-in mechanisms to keep track of individual participants in a conversation. Therefore, it's important to explicitly mention the name or identifier of the participant you are addressing when providing instructions or asking questions.



ChatGPT

ChatGPT's performance on MPC tasks leaves much to be desired, whilst GPT-4's results portend a promising future

Instruction:

*You have been presented with a sequence of multi-party conversational turns, organized in chronological order. Please evaluate the emotions of each utterances in the dialogue using the following 7 labels: {'fear', 'neutral', 'joy', 'disgust', 'anger', 'sadness', 'surprise'}. The output format must be: `#{num} -- {speaker}: {utterance} // {emotion}`
Use temperature=0, minimize unnecessary words to not get confused.*

Input:

#0 -- All: Push, push!
#1 -- Ross: We're here!
#2 -- Carol: Where have you been?
#3 -- Ross: Long story, honey.
#4 -- Dr. Franzblau: All right, Carol, I need you to keep pushing. I need — Excuse me, could I have this?

Response:

#0 -- All: Push, push! // neutral
#1 -- Ross: We're here! // joy
#2 -- Carol: Where have you been? // neutral
#3 -- Ross: Long story, honey. // neutral
#4 -- Dr. Franzblau: All right, Carol, I need you to keep pushing. I need — Excuse me, could I have this? // neutral

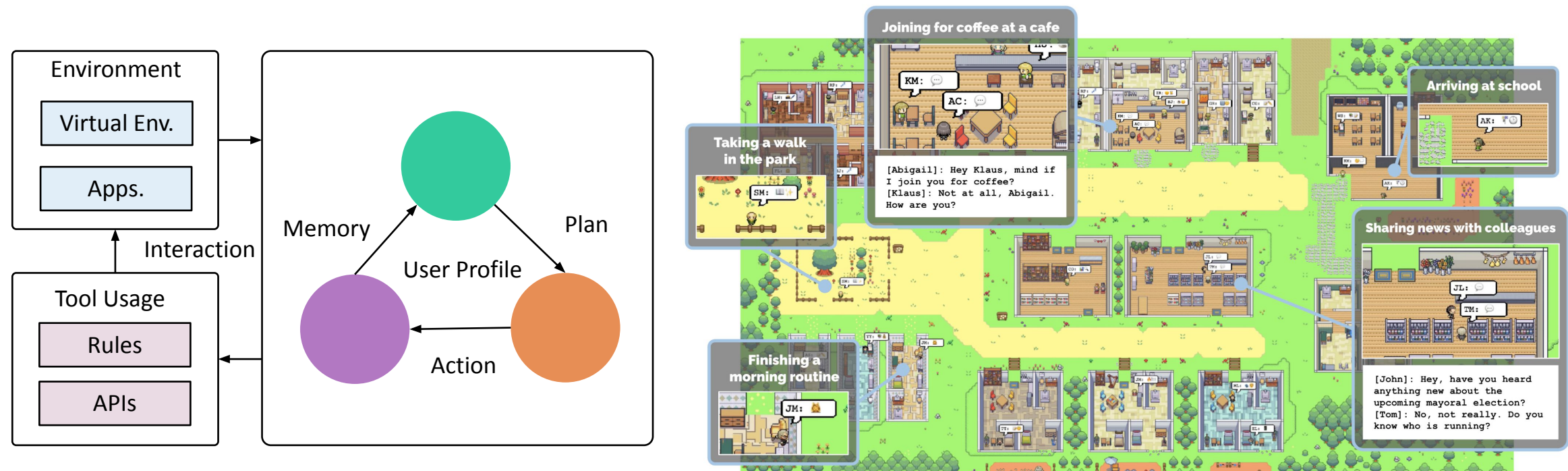
Challenge (3): Multi-modal MPC

Incorporate the face and speech interactions in conversations



Challenge (4): Virtual Agent Simulacra

Instantiate generative agents and empower multi-party interactions to simulate believable human behavior, e.g., collaborate, debate, and evolve



MPC in a Multi-Agent Perspective

Open Questions

- Still don't know yet how to better model the core issues of **interlocutor and conversation structure**
- How to make use of the **complicated** and **complementary** interactions between interlocutors and utterances for **universal** MPC understanding
- Why performance drops significantly when it comes to **topic transition** and **long context**
- How to design **communitive agents**: multimodal environment interaction, multi-agent collaboration
- Lack of high-quality MPC **datasets**



Q & A

Thank you for joining us today!

All the materials are at

http://home.ustc.edu.cn/~gujc/presentations/2023_AACL_MPC.pdf

