Multi-Party Conversation Understanding and Generation

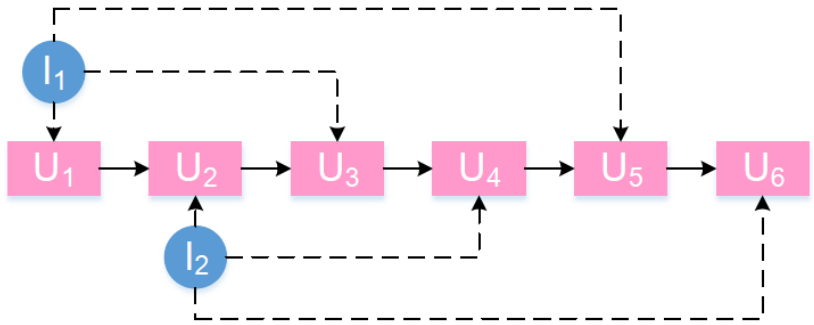
Jia-Chen Gu

National Engineering Research Center for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China

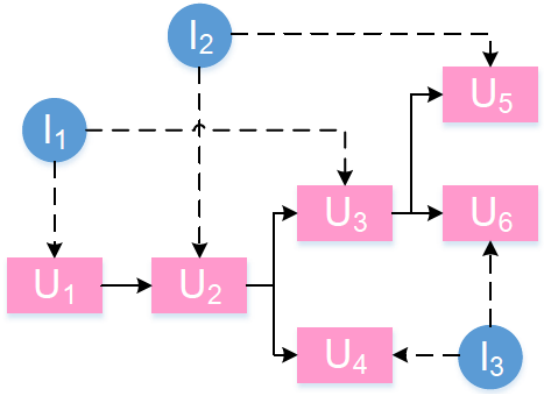
Outline

- **Introduction**
- MPC-BERT
- HeterMPC
- Conclusion

Introduction



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 (MPC) can be spoken by anyone and address anyone else, constituting a graphical information flow.

: Interlocutors

: Utterances

Related Work

- Pre-trained language models still overlook the inherent relationships between utterances and interlocutors, such as “address-to”.
- Existing studies design models for each individual task in MPC separately (e.g., *who says*, *say what* and *address whom*), while neglect the complementary information among these tasks.
- Model a conversation with a homogeneous graph, where nodes represented only utterances while interlocutors are ignored.
- The same forward and backward message passing algorithm cannot distinguish the bidirectional relations between two connected nodes.

Motivation

- Can a language model be pre-trained towards universal MPC understanding?
  - ✓Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, et al. 2021. *MPC-BERT: A pre-trained language model for multi-party conversation understanding*. In *Proc. ACL*, pages 3682–3692.

- Can an MPC be modeled as a heterogeneous graph to embrace various sources of information?
  - ✓Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, et al. 2022. *HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations*. In *Proc. ACL*, pages 5086–5097.

  - ✓Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling. 2022. *Who Says What to Whom: A Survey of Multi-Party Conversations*. In *Proc. IJCAI*, pages 5486–5493.

Outline

- Introduction
- **MPC-BERT**
- HeterMPC
- Conclusion

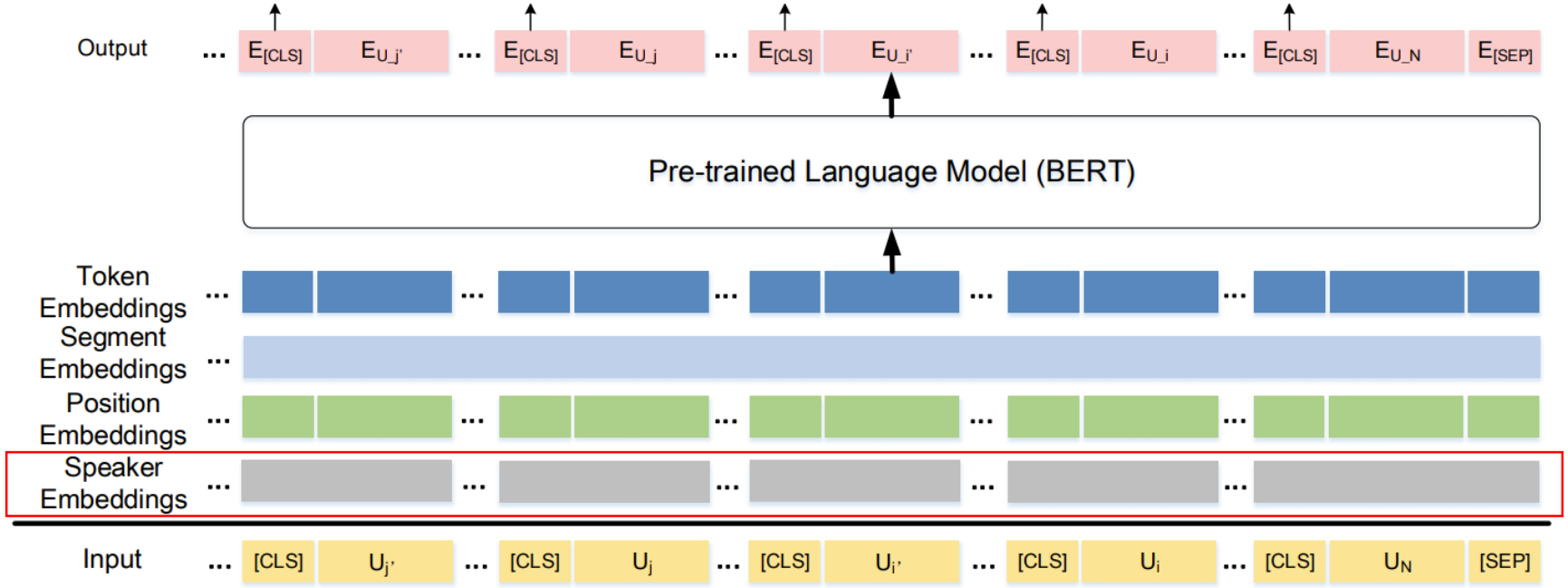
MPC-BERT

The goal is to pre-train language models for universal MPC understanding. MPC-BERT jointly learns ***who says what to whom*** in MPC by designing self-supervised tasks, so that it can produce better interlocutor and utterance representations which can be effectively generalized to multiple downstream tasks of MPC.

- Interlocutor Structure Modeling
- Utterance Semantics Modeling

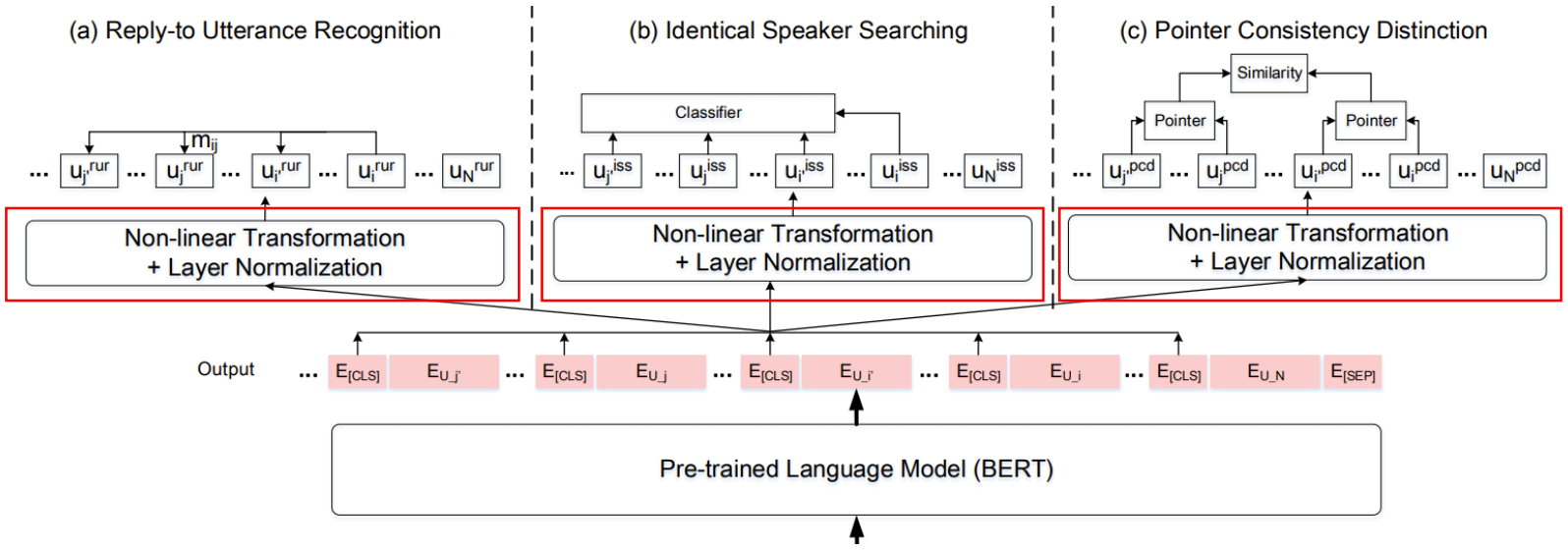
Model overview of MPC-BERT

- A [CLS] token is inserted at the start of each utterance.
- Position-based speaker embeddings are introduced considering that the set of interlocutors are inconsistent in different conversations.



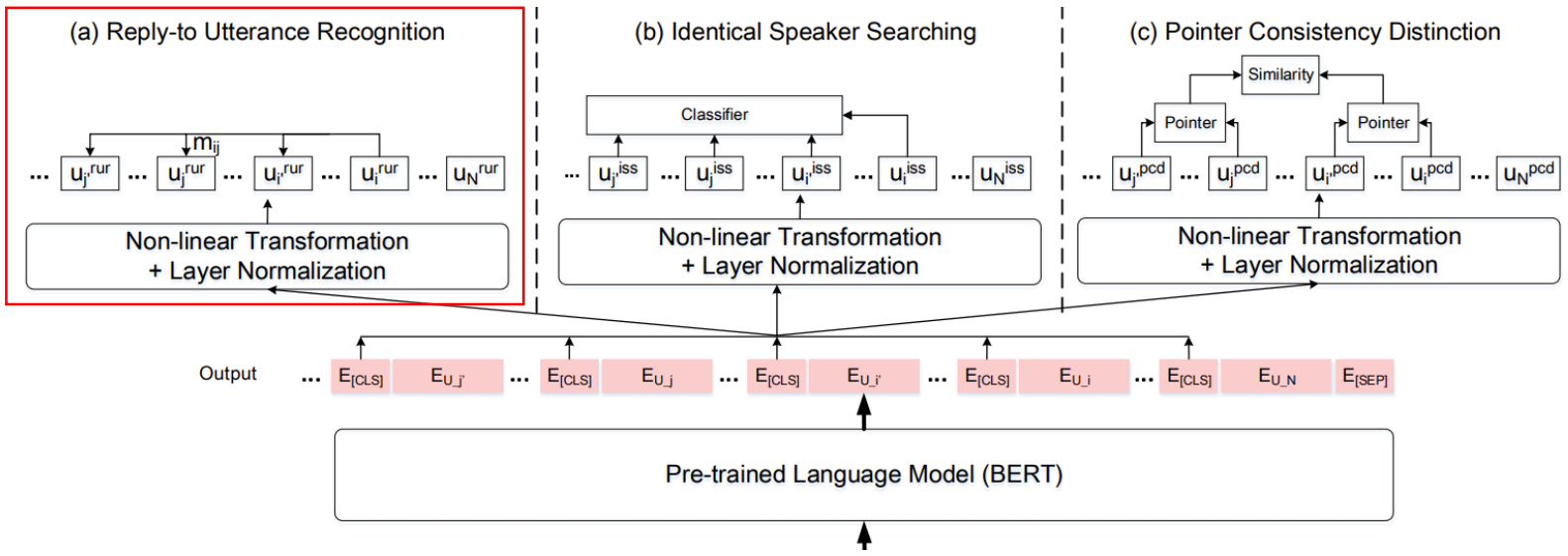
Interlocutor Structure Modeling

- Extract the contextualized representations for each [CLS] token representing individual utterances.
- A task-dependent non-linear transformation is placed on top of BERT.
- Encoding the input data only once is computation-efficient.



Interlocutor Structure Modeling

- **Reply-to Utterance Recognition**: To enable the model to recognize the addressee of each utterance, this task is proposed to learn which preceding utterance the current utterance replies to.

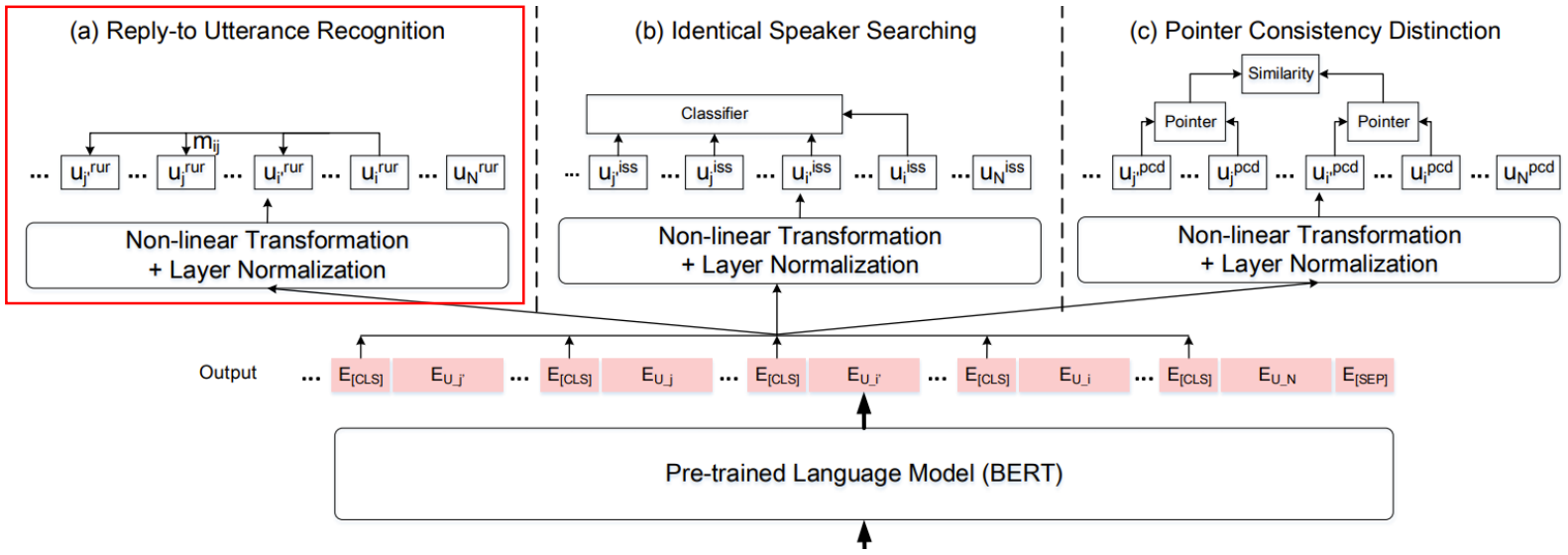


Interlocutor Structure Modeling

- **Reply-to Utterance Recognition**: For a specific utterance U_i, its matching scores with all its preceding utterances are calculated as

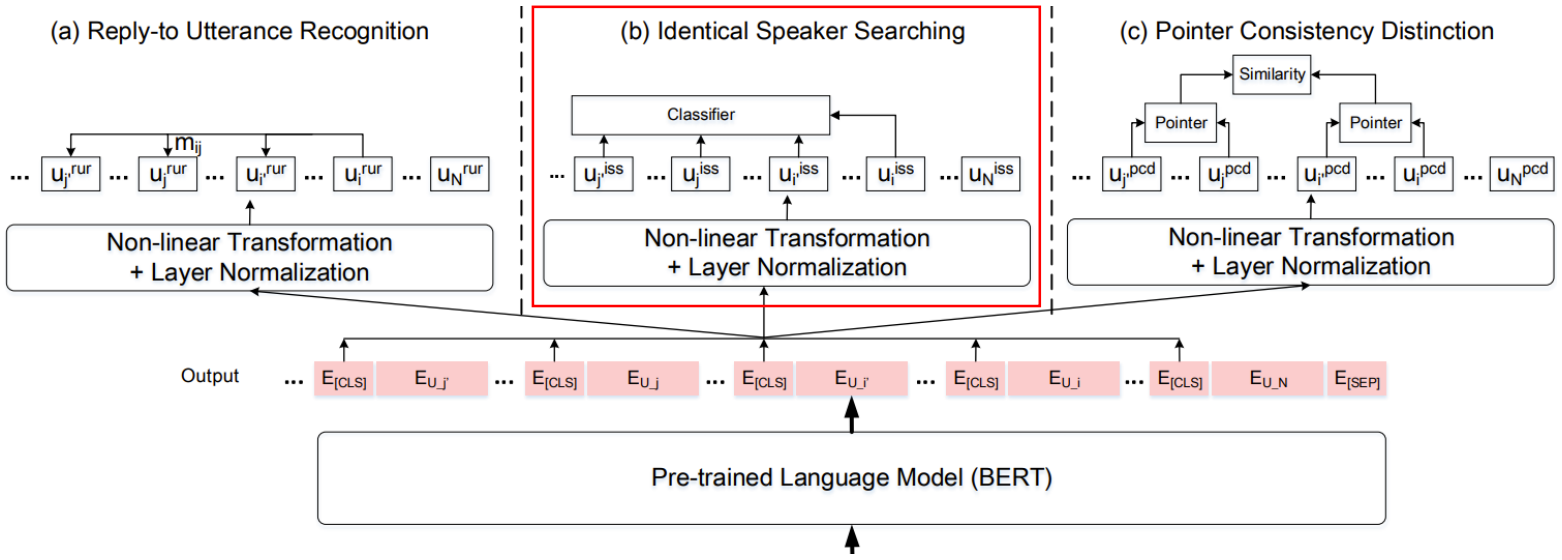
$$m_{ij} = \mathbf{softmax}(\mathbf{u}_i^{rur\top} \cdot \mathbf{A}^{rur} \cdot \mathbf{u}_j^{rur})$$

- Dynamic sampling + Cross-entropy loss minimization $\mathcal{L}_{rur} = -\sum_{i \in \mathbb{S}} \sum_{j=1}^{i-1} y_{ij} \, log(m_{ij})$



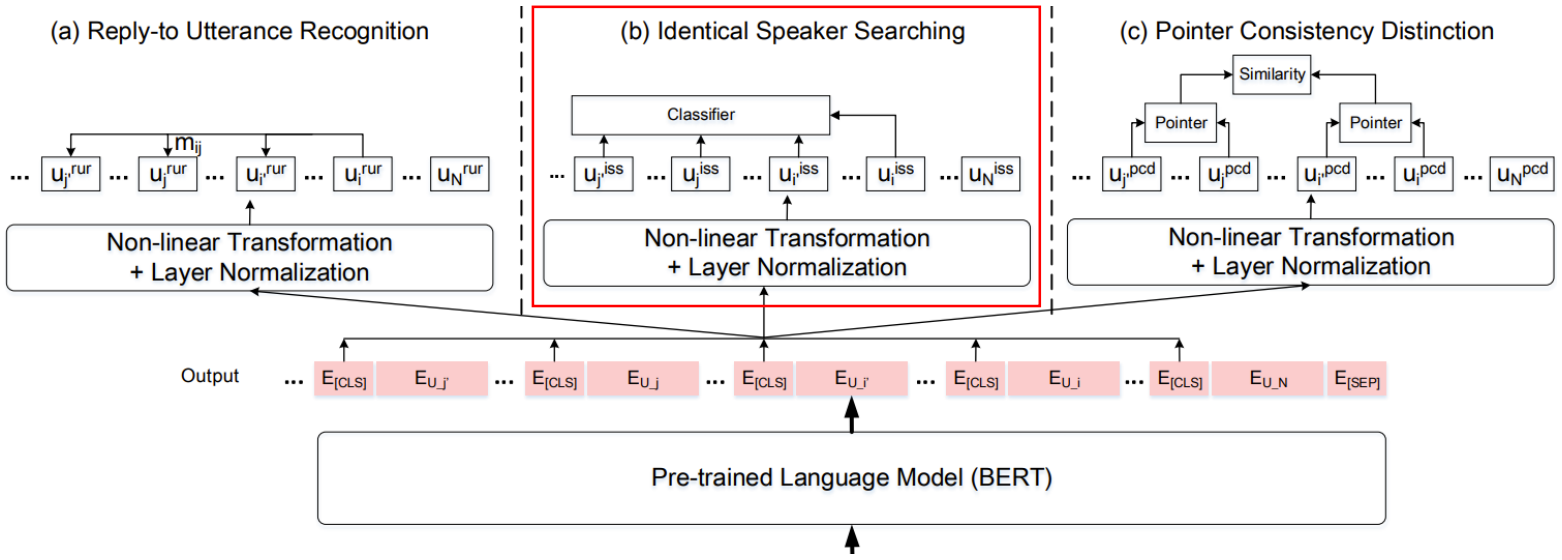
Interlocutor Structure Modeling

- **Identical Speaker Searching:** Since the set of interlocutors vary across conversations, the task of predicting the speaker of an utterance is reformulated as searching for the utterances sharing the identical speaker.



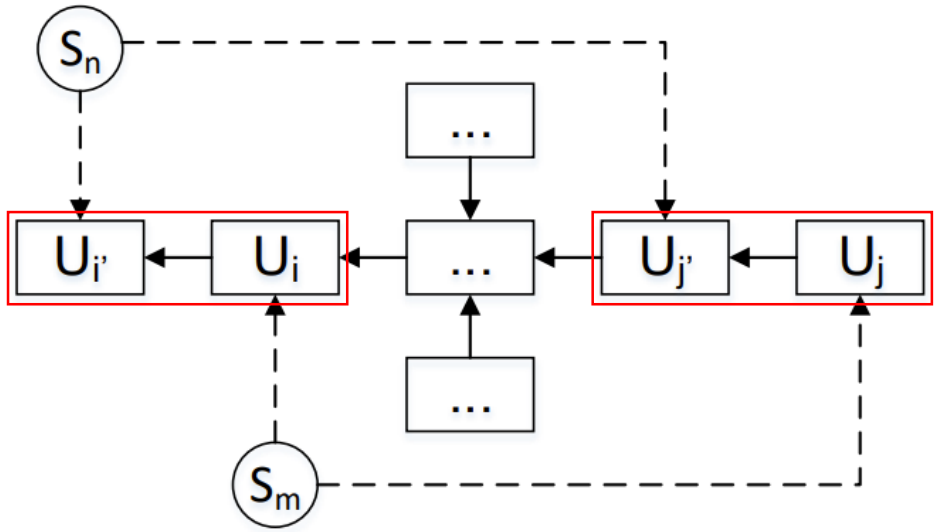
Interlocutor Structure Modeling

- **Identical Speaker Searching:** Mask the speaker embedding of a specific utterance in the input representation, and calculate the probability of two utterances sharing the same speaker.
- Dynamic sampling + Cross-entropy loss minimization



Interlocutor Structure Modeling

- **Pointer Consistency Distinction:** A pair of utterances representing the "reply-to" relationship is defined as a speaker-to-addressee pointer.
- We assume that the representations of two pointers directing from the same speaker to the same addressee should be consistent.

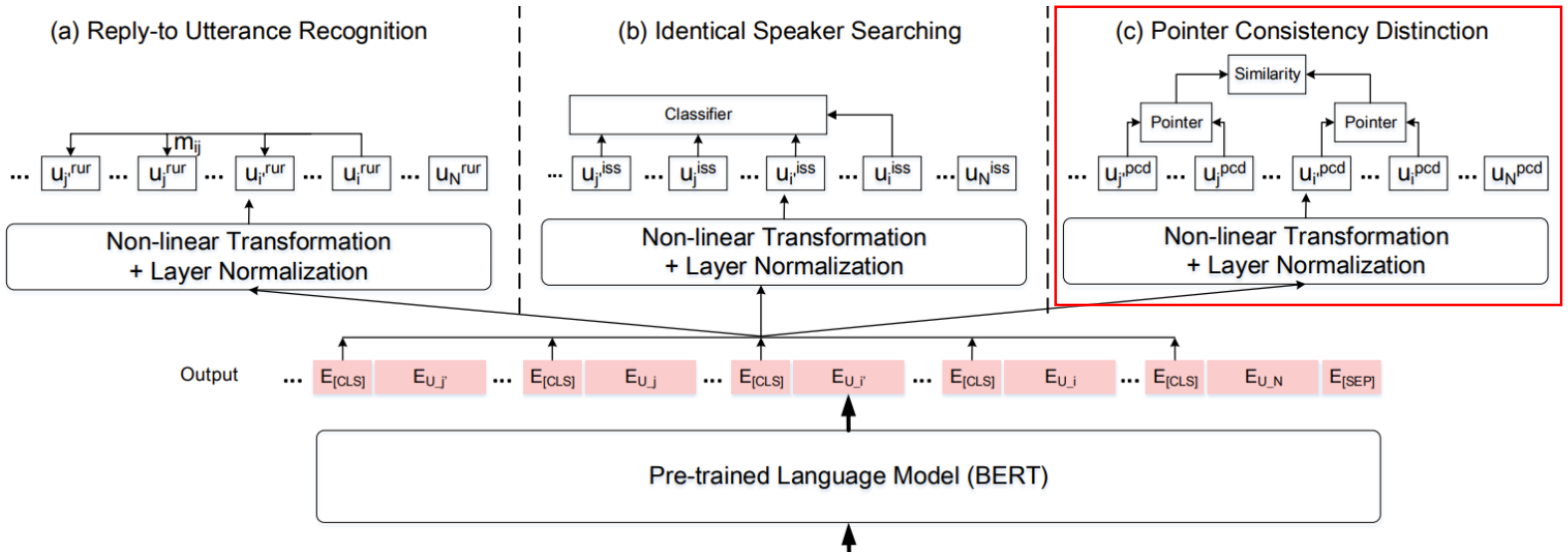


Interlocutor Structure Modeling

- **Pointer Consistency Distinction:** Capture the pointer information contained in each utterance tuple as

$$\mathbf{p}_{ii'} = [\mathbf{u}_i^{pcd} - \mathbf{u}_{i'}^{pcd}; \mathbf{u}_i^{pcd} \odot \mathbf{u}_{i'}^{pcd}]$$

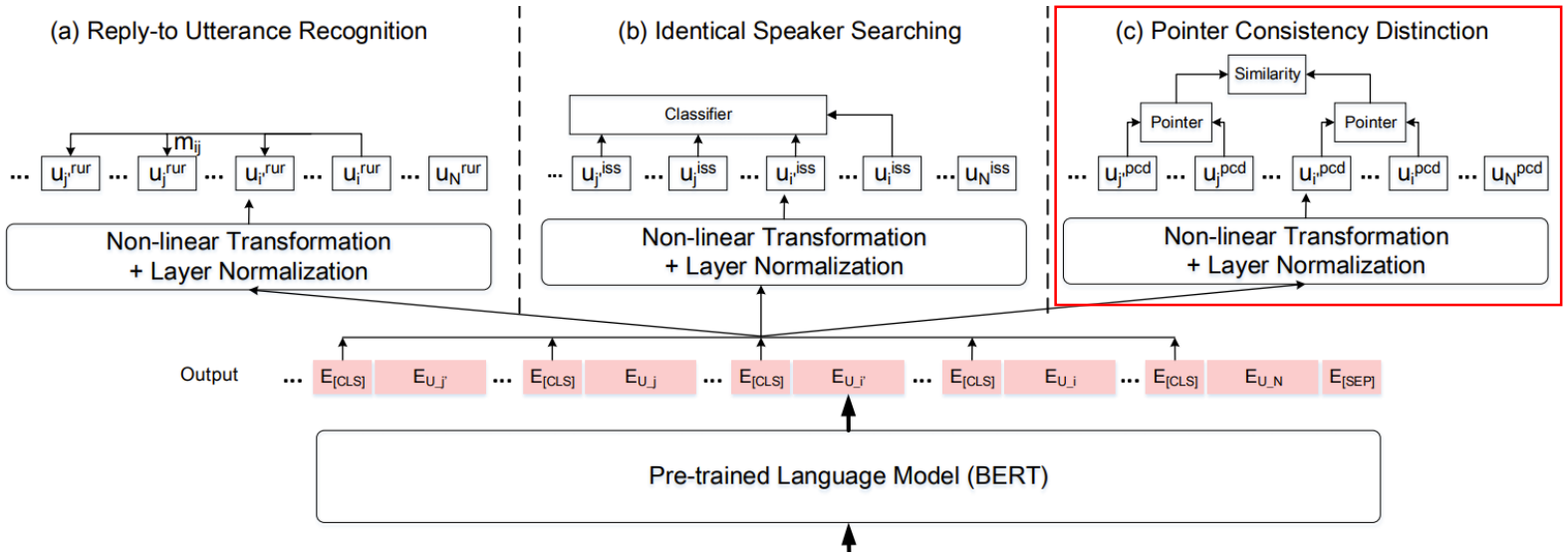
$$\bar{\mathbf{p}}_{ii'} = \mathbf{ReLU}(\mathbf{p}_{ii'} \cdot \mathbf{W}_{pcd} + \mathbf{b}_{pcd})$$



Interlocutor Structure Modeling

- **Pointer Consistency Distinction:** A consistent pointer representations and an inconsistent one sampled from this conversation are obtained. The similarities between every two pointers are calculated as

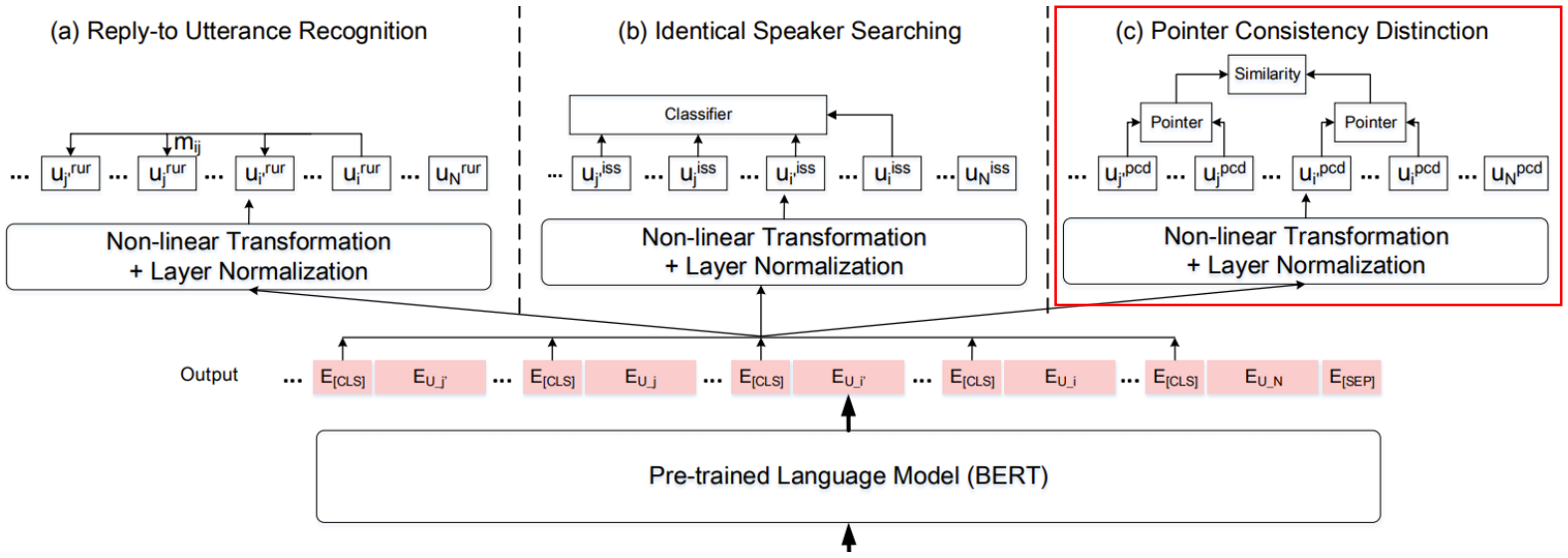
$$m_{ij} = \mathbf{sigmoid}(\bar{\mathbf{p}}_{ii'}^{\top} \cdot \mathbf{A}^{pcd} \cdot \bar{\mathbf{p}}_{jj'})$$



Interlocutor Structure Modeling

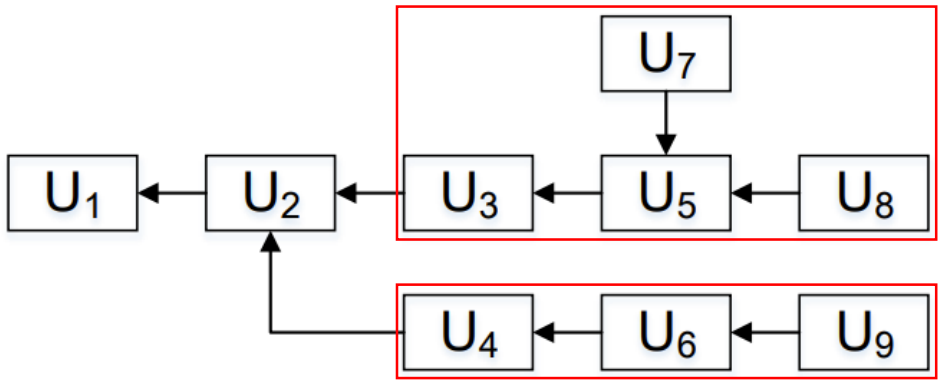
- **Pointer Consistency Distinction:** Minimize the hinge loss which enforces m_{ij} to be larger than m_{ik} by at least a margin Δ as

$$\mathcal{L}_{pcd} = \mathbf{max}\{0, \Delta - m_{ij} + m_{ik}\}$$



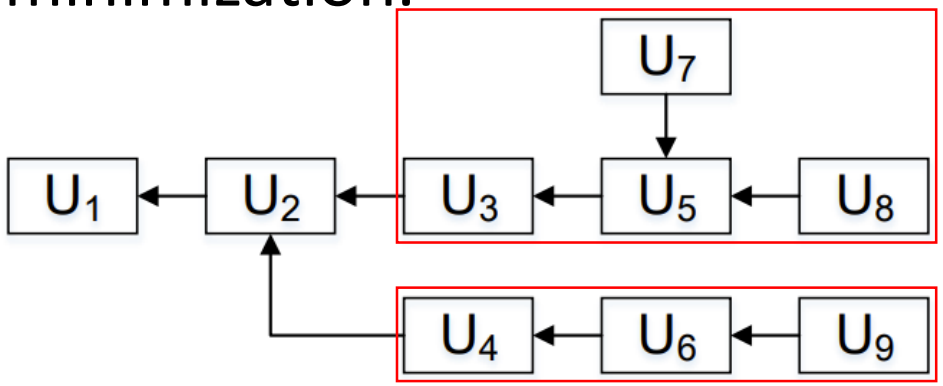
Utterance Semantics Modeling

- **Shared Node Detection:** A full MPC instance can be divided into several sub-conversations and we assume that the representations of sub-conversations under the same parent node tend to be similar.
- For example, two sub-conversations {U3, U5, U7, U8} and {U4, U6, U9} share the same parent node U2.



Utterance Semantics Modeling

- **Shared Node Detection:** Given a full MPC, the two sub-conversations under the top shared node (most utterances) form a positive pair empirically. Replace one sub-conversation with another one randomly sampled from the training corpus to form a negative pair.
- Sequence-pair prediction with the representation of the [CLS] token.
- Cross-entropy loss minimization.



Utterance Semantics Modeling

- **Masked Shared Utterance Restoration:** There are usually several utterances replying-to a shared utterance in MPC. A shared utterance is semantically relevant to more utterances in the context than non-shared ones.
- All tokens in a sampled shared utterance are masked with a [MASK] token and the model is enforced to restore the masked utterance given the rest conversation. (Utterance-level Language Model)

Multi-task Learning

- The tasks of masked language model (MLM) and next sentence prediction (NSP) in original BERT pre-training are also adopted, which have been proven effective for incorporating domain knowledge.
- MPC-BERT is trained by performing multi-task learning that minimizes the sum of all loss functions as

$$\mathcal{L} = \mathcal{L}_{rur} + \mathcal{L}_{iss} + \mathcal{L}_{pcd} + \mathcal{L}_{msur} + \mathcal{L}_{snd} + \mathcal{L}_{mlm} + \mathcal{L}_{nsp}$$

Downstream Tasks

- To measure the effectiveness of these self-supervised tasks and to test the generalization ability of MPC-BERT, we evaluate MPC-BERT on three downstream tasks including addressee recognition, speaker identification and response selection, which are three core research issues of MPC.

Addressee Recognition

- In this paper, we follow the more challenging setting in Le et al. (2019) where addressees of all utterances in a conversation are asked to recognized.
- Given $\{(s_n, u_n, a_n)\}_{n=1}^{N} \setminus \{a_n\}_{n=1}^{N}$, models are asked to predict $\{\hat{a}_n\}_{n=1}^{N}$ where $\hat{a}_n$ is selected from the interlocutor set in this conversation.

*a, u, s and / denote addressee, utterance, speaker and exclusion respectively.

Speaker Identification

- This task aims to identify the speaker of the last utterance in a conversation, where the identified speaker is selected from the interlocutor set in this conversation.
- Given $\{(s_n, u_n, a_n)\}_{n=1}^{N} \setminus s_N$, models are asked to predict $\hat{s}_N$, where $\hat{s}_N$ is selected from the interlocutor set in this conversation.

Response Selection

- This tasks aims to measure the similarity between a context and a response, and then rank a set of response candidates, which is an important retrieval-based approach for chatbots.
- This task asks models to select $\hat{u}_N$ from a set of response candidates given the conversation context $\{(s_n, u_n, a_n)\}_{n=1}^{N} \setminus u_N$.

Experiments

- Datasets

  We evaluated MPC-BERT on two Ubuntu IRC benchmarks.

Datasets		Train	Valid	Test
Hu et al. (2019)		311,725	5,000	5,000
Ouchi and Tsuboi (2016)	Len-5	461,120	28,570	32,668
	Len-10	495,226	30,974	35,638
	Len-15	489,812	30,815	35,385

Addressee Recognition

- Precision@1 (P@1) to evaluate each utterance with ground truth. Accuracy (Acc.) to evaluate a session if all addressees are recognized.

	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	P@1	Acc.	P@1	Acc.	P@1	Acc.	P@1	Acc.
Preceding (Le et al., 2019)	-	-	63.50	40.46	56.84	21.06	54.97	13.08
Subsequent (Le et al., 2019)	-	-	61.03	40.25	54.57	20.26	53.07	12.79
DRNN (Ouchi and Tsuboi, 2016)	-	-	72.75	58.18	65.58	34.47	62.60	22.58
SIRNN (Zhang et al., 2018)	-	-	75.98	62.06	70.88	40.66	68.13	28.05
W2W (Le et al., 2019)	-	-	77.55	63.81	73.52	44.14	73.42	34.23
BERT (Devlin et al., 2019)	96.16	83.50	85.95	75.99	83.41	58.22	81.09	44.94
SA-BERT (Gu et al., 2020a)	97.12	88.91	86.81	77.45	84.46	60.30	82.84	47.23
MPC-BERT	**98.31**	**92.42**	**88.73**	**80.31**	**86.23**	**63.58**	**85.55**	**52.59**
MPC-BERT w/o. RUR	97.75	89.98	87.51	78.42	85.63	62.26	84.78	50.83
MPC-BERT w/o. ISS	98.20	91.96	88.67	80.25	86.14	63.40	85.02	51.12
MPC-BERT w/o. PCD	98.20	91.90	88.51	80.06	85.92	62.84	85.21	51.17
MPC-BERT w/o. MSUR	98.08	91.32	88.70	80.26	86.21	63.46	85.28	51.23
MPC-BERT w/o. SND	98.25	92.18	88.68	80.25	86.14	63.41	85.29	51.39

Table 3: Evaluation results of addressee recognition on the test sets. Results except ours are cited from Le et al. (2019). Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05).

Addressee Recognition

- MPC-BERT outperforms SA-BERT by margins of 3.51%, 2.86%, 3.28% and 5.36% on these test sets respectively in terms of Acc.

	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	P@1	Acc.	P@1	Acc.	P@1	Acc.	P@1	Acc.
Preceding (Le et al., 2019)	-	-	63.50	40.46	56.84	21.06	54.97	13.08
Subsequent (Le et al., 2019)	-	-	61.03	40.25	54.57	20.26	53.07	12.79
DRNN (Ouchi and Tsuboi, 2016)	-	-	72.75	58.18	65.58	34.47	62.60	22.58
SIRNN (Zhang et al., 2018)	-	-	75.98	62.06	70.88	40.66	68.13	28.05
W2W (Le et al., 2019)	-	-	77.55	63.81	73.52	44.14	73.42	34.23
BERT (Devlin et al., 2019)	96.16	83.50	85.95	75.99	83.41	58.22	81.09	44.94
SA-BERT (Gu et al., 2020a)	97.12	88.91	86.81	77.45	84.46	60.30	82.84	47.23
MPC-BERT	**98.31**	**92.42**	**88.73**	**80.31**	**86.23**	**63.58**	**85.55**	**52.59**
MPC-BERT w/o. RUR	97.75	89.98	87.51	78.42	85.63	62.26	84.78	50.83
MPC-BERT w/o. ISS	98.20	91.96	88.67	80.25	86.14	63.40	85.02	51.12
MPC-BERT w/o. PCD	98.20	91.90	88.51	80.06	85.92	62.84	85.21	51.17
MPC-BERT w/o. MSUR	98.08	91.32	88.70	80.26	86.21	63.46	85.28	51.23
MPC-BERT w/o. SND	98.25	92.18	88.68	80.25	86.14	63.41	85.29	51.39

Table 3: Evaluation results of addressee recognition on the test sets. Results except ours are cited from Le et al. (2019). Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05).

Addressee Recognition

- RUR contributes the most, and the tasks modeling interlocutor structure contribute more than those for utterance semantics.

	Hu et al. (2019)		Ouchi and Tsuboi (2016)					
			Len-5		Len-10		Len-15	
	P@1	Acc.	P@1	Acc.	P@1	Acc.	P@1	Acc.
Preceding (Le et al., 2019)	-	-	63.50	40.46	56.84	21.06	54.97	13.08
Subsequent (Le et al., 2019)	-	-	61.03	40.25	54.57	20.26	53.07	12.79
DRNN (Ouchi and Tsuboi, 2016)	-	-	72.75	58.18	65.58	34.47	62.60	22.58
SIRNN (Zhang et al., 2018)	-	-	75.98	62.06	70.88	40.66	68.13	28.05
W2W (Le et al., 2019)	-	-	77.55	63.81	73.52	44.14	73.42	34.23
BERT (Devlin et al., 2019)	96.16	83.50	85.95	75.99	83.41	58.22	81.09	44.94
SA-BERT (Gu et al., 2020a)	97.12	88.91	86.81	77.45	84.46	60.30	82.84	47.23
MPC-BERT	**98.31**	**92.42**	**88.73**	**80.31**	**86.23**	**63.58**	**85.55**	**52.59**
MPC-BERT w/o. RUR	97.75	89.98	87.51	78.42	85.63	62.26	84.78	50.83
MPC-BERT w/o. ISS	98.20	91.96	88.67	80.25	86.14	63.40	85.02	51.12
MPC-BERT w/o. PCD	98.20	91.90	88.51	80.06	85.92	62.84	85.21	51.17
MPC-BERT w/o. MSUR	98.08	91.32	88.70	80.26	86.21	63.46	85.28	51.23
MPC-BERT w/o. SND	98.25	92.18	88.68	80.25	86.14	63.41	85.29	51.39

Table 3: Evaluation results of addressee recognition on the test sets. Results except ours are cited from Le et al. (2019). Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05).

Speaker Identification

- Precision@1 (P@1) to evaluate the last utterance of a conversation.
- MPC-BERT outperforms SA-BERT by margins of 7.66%, 2.60%, 3.38% and 4.24% respectively in terms of P@1.
- 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).

Response Selection

- R_n@k to evaluate top-*k* selected responses from *n* available candidates. Two settings of R_2@1 and R_{10}@1 were followed.

	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., 2020a)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT	**94.90**	**78.98**	**87.63**	**57.95**	**89.14**	**61.82**	**89.70**	**63.64**
MPC-BERT w/o. RUR	94.48	78.16	87.20	57.56	88.96	61.47	89.07	63.24
MPC-BERT w/o. ISS	94.58	78.82	87.54	57.77	88.98	61.76	89.58	63.51
MPC-BERT w/o. PCD	94.66	78.70	87.50	57.51	88.75	61.62	89.45	63.46
MPC-BERT w/o. MSUR	94.36	78.22	87.11	57.58	88.59	61.05	89.25	63.20
MPC-BERT w/o. SND	93.92	76.96	87.30	57.54	88.77	61.54	89.27	63.34

Table 5: Evaluation results of response selection on the test sets. Results except ours are cited from Ouchi and Tsuboi (2016) and Zhang et al. (2018). Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05).

Response Selection

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

	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., 2020a)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT	**94.90**	**78.98**	**87.63**	**57.95**	**89.14**	**61.82**	**89.70**	**63.64**
MPC-BERT w/o. RUR	94.48	78.16	87.20	57.56	88.96	61.47	89.07	63.24
MPC-BERT w/o. ISS	94.58	78.82	87.54	57.77	88.98	61.76	89.58	63.51
MPC-BERT w/o. PCD	94.66	78.70	87.50	57.51	88.75	61.62	89.45	63.46
MPC-BERT w/o. MSUR	94.36	78.22	87.11	57.58	88.59	61.05	89.25	63.20
MPC-BERT w/o. SND	93.92	76.96	87.30	57.54	88.77	61.54	89.27	63.34

Table 5: Evaluation results of response selection on the test sets. Results except ours are cited from Ouchi and Tsuboi (2016) and Zhang et al. (2018). Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05).

Response Selection

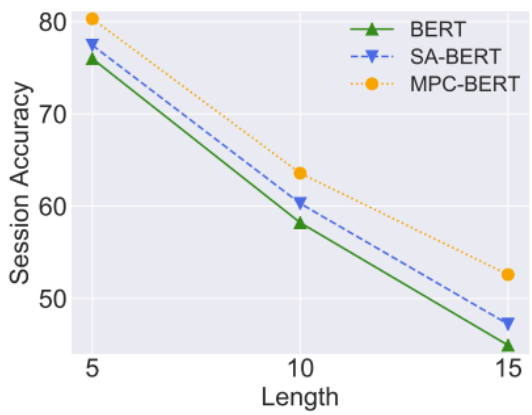
- SND contributes the most, and the two tasks modeling the utterance semantics contribute more than those for the interlocutor structures.

	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., 2020a)	92.98	75.16	86.53	55.24	87.98	59.27	88.34	60.42
MPC-BERT	**94.90**	**78.98**	**87.63**	**57.95**	**89.14**	**61.82**	**89.70**	**63.64**
MPC-BERT w/o. RUR	94.48	78.16	87.20	57.56	88.96	61.47	89.07	63.24
MPC-BERT w/o. ISS	94.58	78.82	87.54	57.77	88.98	61.76	89.58	63.51
MPC-BERT w/o. PCD	94.66	78.70	87.50	57.51	88.75	61.62	89.45	63.46
MPC-BERT w/o. MSUR	94.36	78.22	87.11	57.58	88.59	61.05	89.25	63.20
MPC-BERT w/o. SND	93.92	76.96	87.30	57.54	88.77	61.54	89.27	63.34

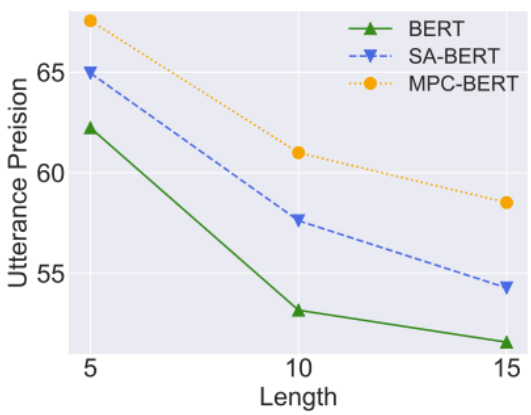
Table 5: Evaluation results of response selection on the test sets. Results except ours are cited from Ouchi and Tsuboi (2016) and Zhang et al. (2018). Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05).

Discussions

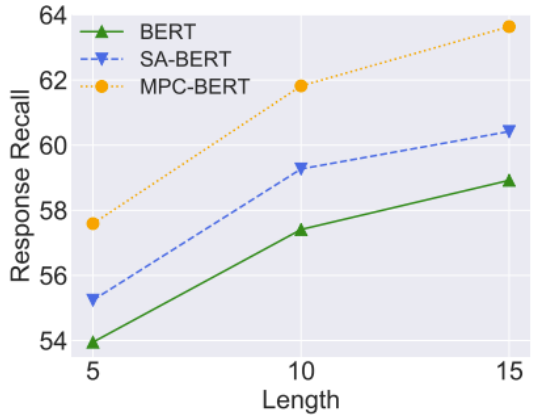
- How the performance of BERT, SA-BERT and MPC-BERT changed with respect to different session lengths on the test sets of Ouchi and Tsuboi (2016).



(a) Addressee recognition



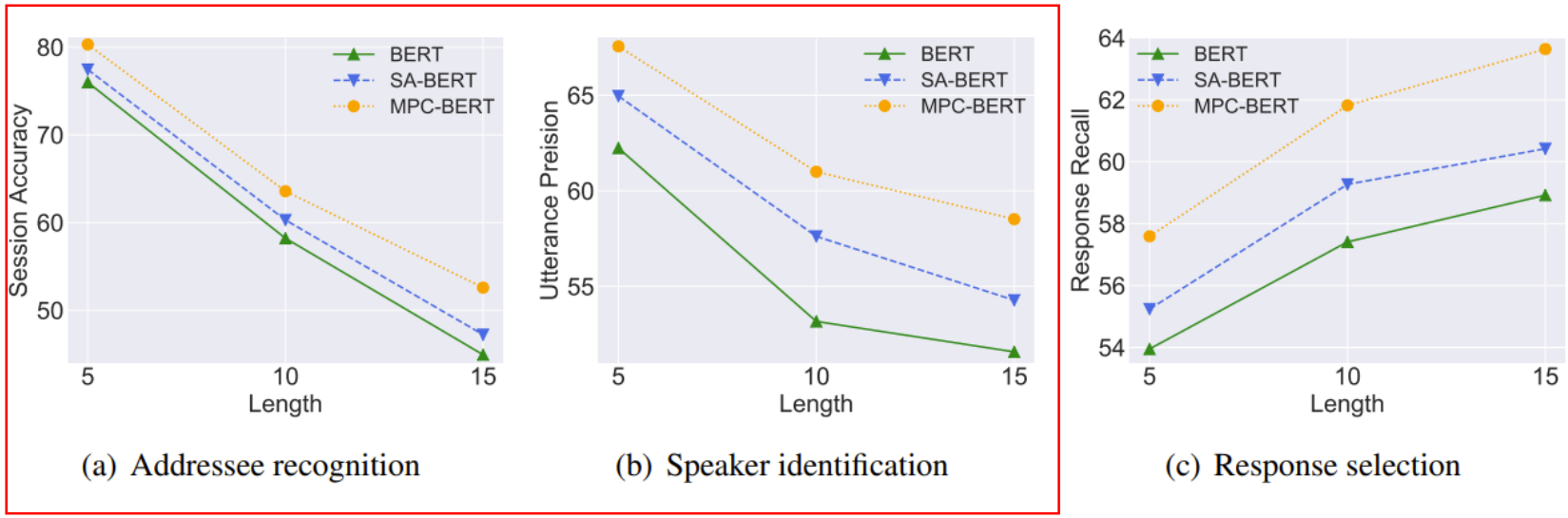
(b) Speaker identification



(c) Response selection

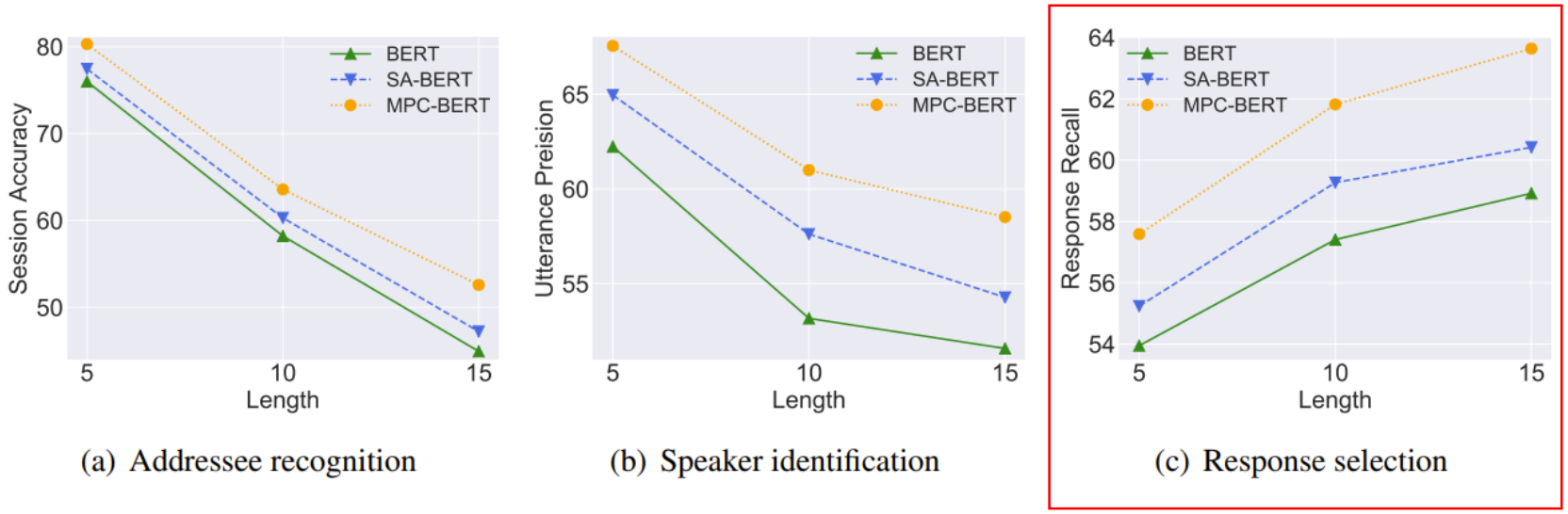
Discussions

- The performance of addressee recognition and speaker identification dropped as the session length increased.
- The reason might be that longer sessions always contain more interlocutors which increase the difficulties of predicting interlocutors.



Discussions

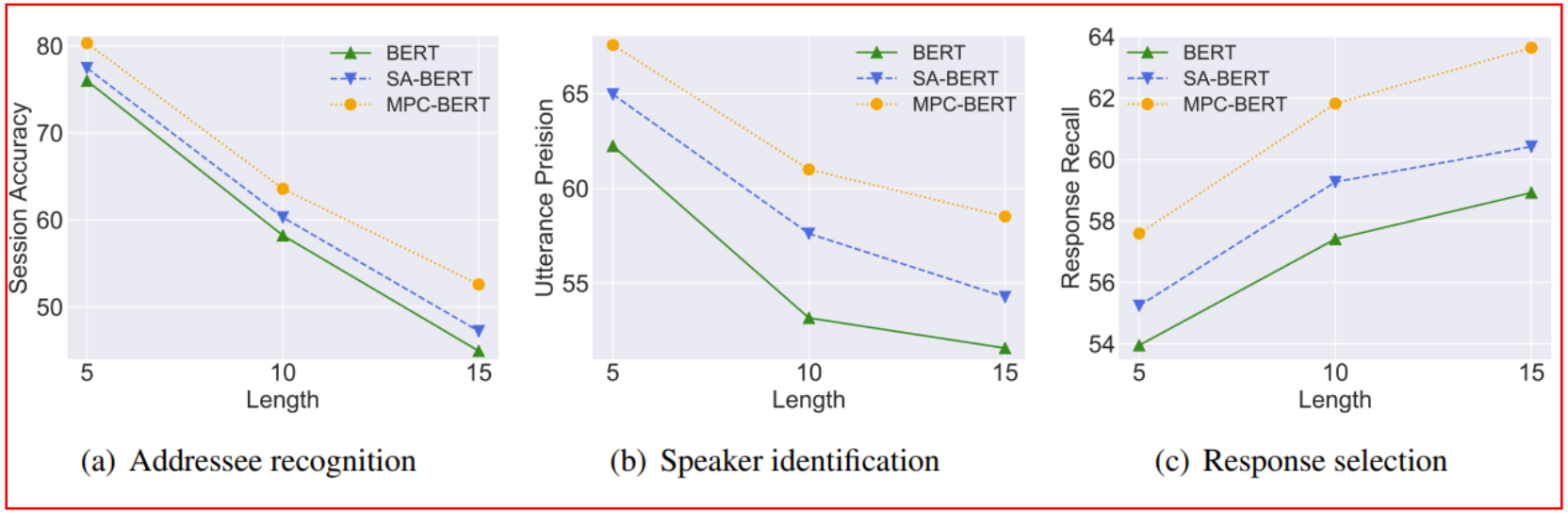
- The performance of response selection was significantly improved as the session length increased.
- It can be attributed to that longer sessions enrich the representations of contexts with more details which benefit response selection.



(a) Addressee recognition (b) Speaker identification (c) Response selection

Discussions

- As the session length increased, the performance of MPC-BERT dropped more slightly than that of SA-BERT on addressee recognition and speaker identification, and the $R_{10}@1$ gap between MPC-BERT and SA-BERT on response selection enlarged from 2.71% to 3.22%.
- Imply superiorities of MPC-BERT on modeling complicated structures.



Outline

- Introduction
- MPC-BERT
- **HeterMPC**
- Conclusion

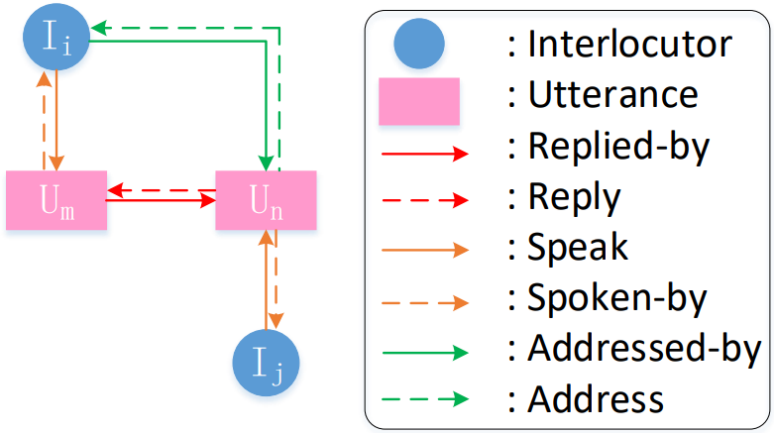
HeterMPC

- Utterances and interlocutors are considered as two types of nodes under a unified heterogeneous graph, to explicitly model the complicated interactions between interlocutors, between utterances, and between an interlocutor and an utterance.

Graph Construction

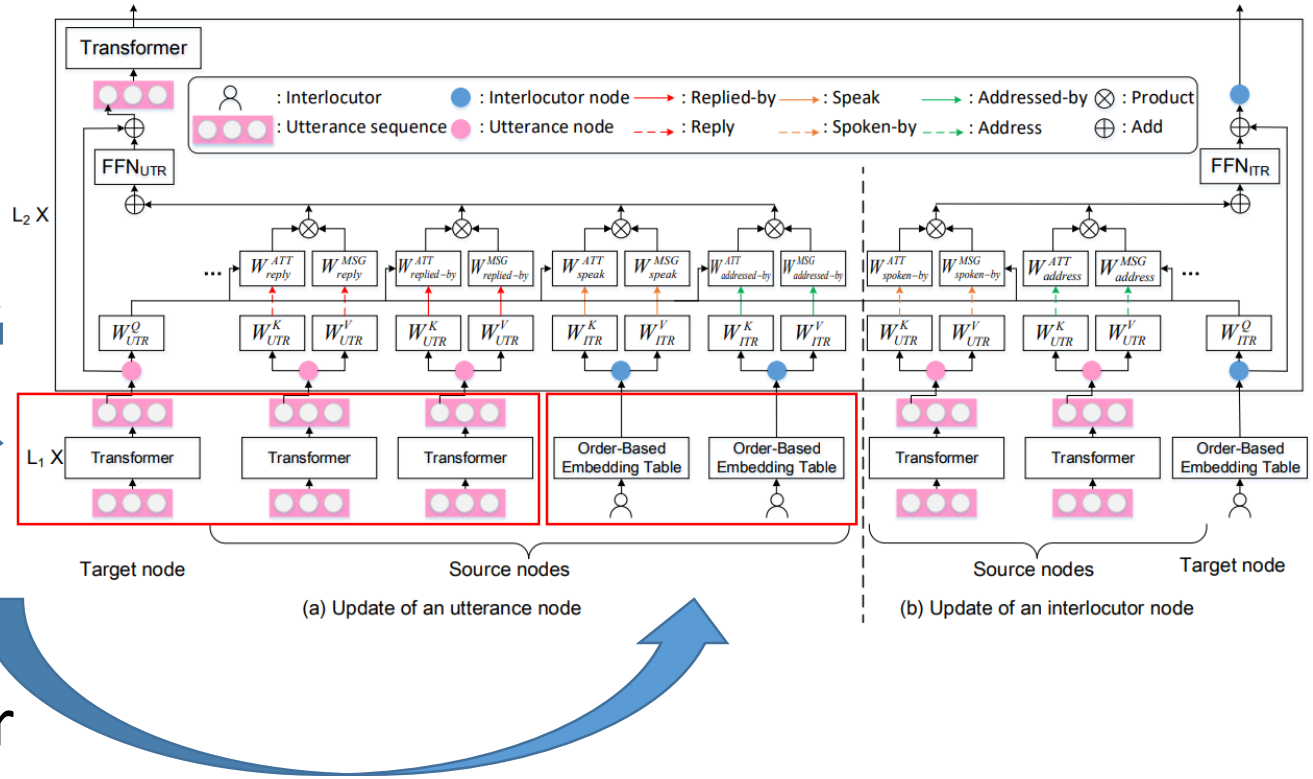
- *M* utterances and *I* interlocutors → a heterogeneous graph *G(V, E)*
- *V* : a set of *M + I* nodes, each denoting an utterance or an interlocutor
- $E = \{e_{p,q}\}_{p,q=1}^{M+I}$: a set of directed edges, each edge $e_{p,q}$ describing the connection from node *p* to node *q*

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



Node Initialization

- Each utterance is encoded individually by stacked Transformer encoder layers
- Each interlocutor is directly represented by looking up an order-based interlocutor embedding table



Node Updating

- Introduce parameters to model heterogeneity
- Attention weights

$$\boldsymbol{k}^l(s) = \boldsymbol{h}_s^l \boldsymbol{W}^K_{\tau(s)} + \boldsymbol{b}^K_{\tau(s)},$$

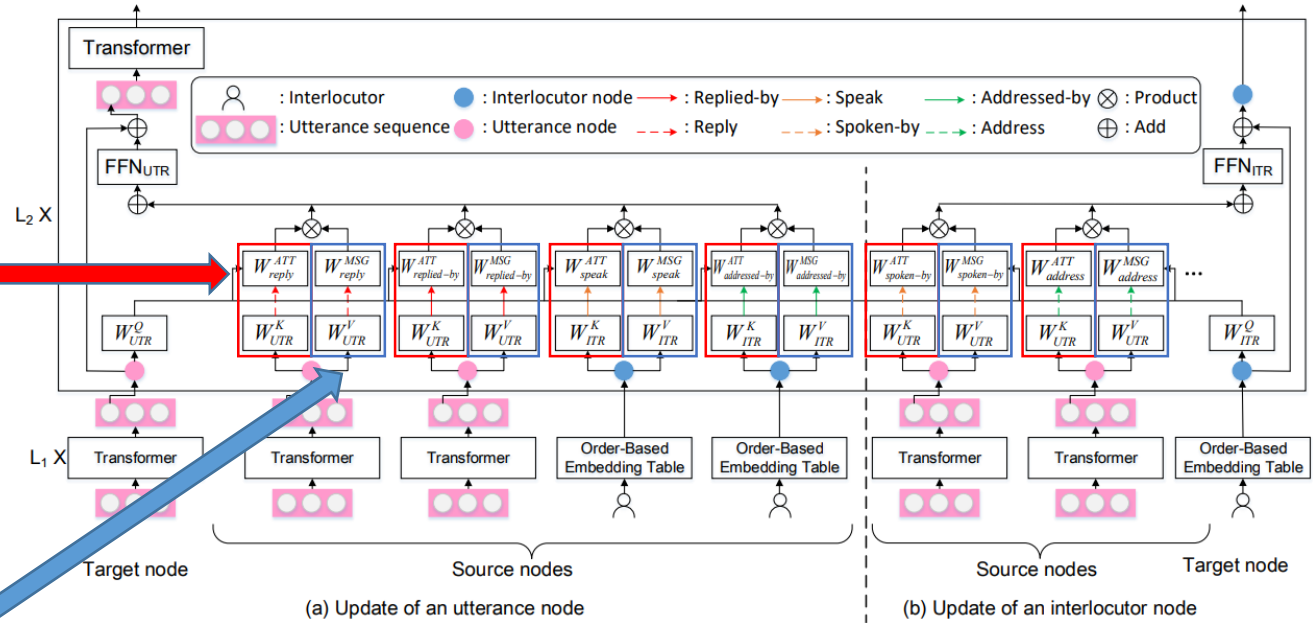
$$\boldsymbol{q}^l(t) = \boldsymbol{h}_t^l \boldsymbol{W}^Q_{\tau(t)} + \boldsymbol{b}^Q_{\tau(t)},$$

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

- Message passing

$$\boldsymbol{v}^l(s) = \boldsymbol{h}_s^l \boldsymbol{W}^V_{\tau(s)} + \boldsymbol{b}^V_{\tau(s)},$$

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



(s, e, t) denotes (source, edge, target)

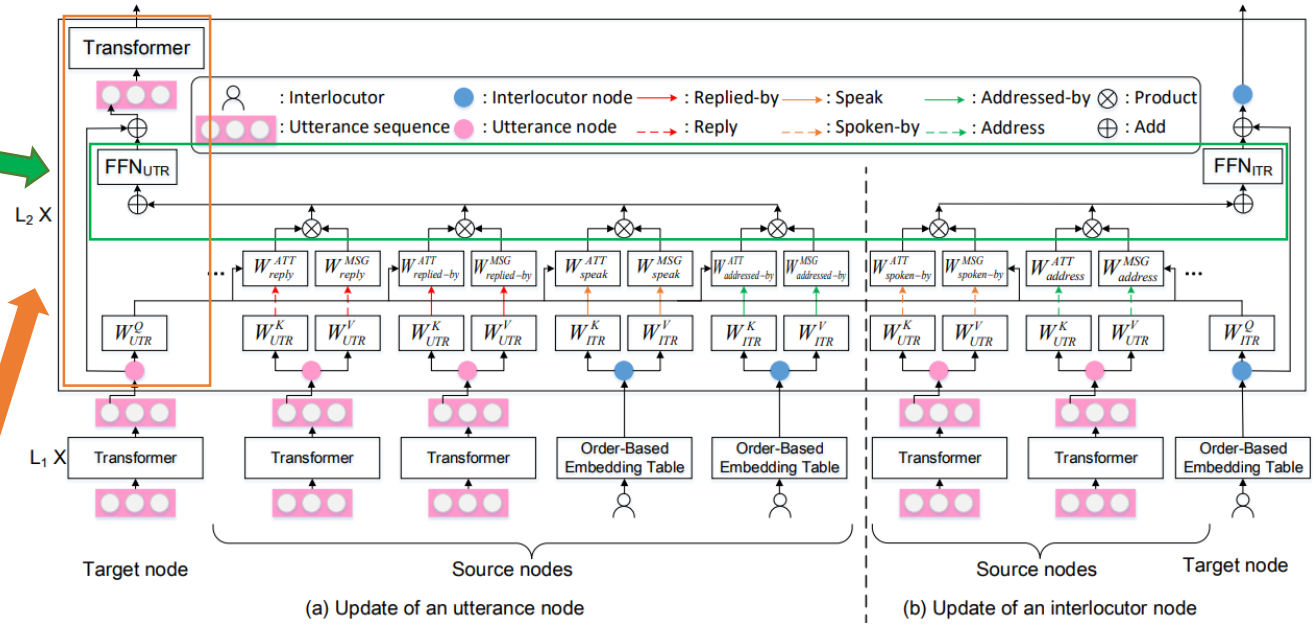
τ(s), τ(t) ∈ {utterance, interlocutor}

Node Updating

- Aggregation

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

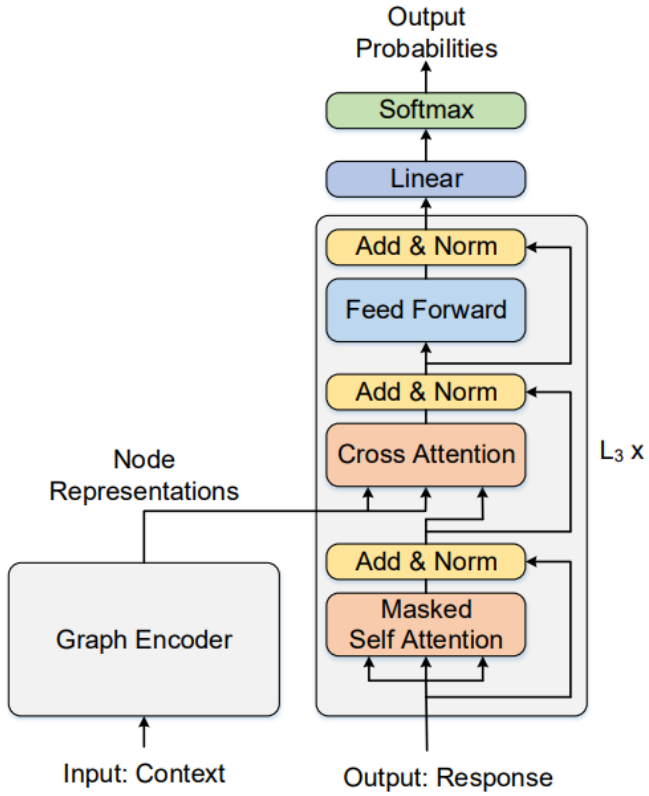
$$\boldsymbol{h}_t^{l+1} = FFN_{\tau(t)}(\bar{\boldsymbol{h}}_t^l) + \boldsymbol{h}_t^l,$$



- Specifically, the context information in an utterance node is shared with other tokens in the utterance through another round of Transformer layer intra-utterance self-attention.

Decoder

- Standard implementation of Transformer decoder
- A cross-attention operation over the node representations of the graph encoder output is performed to incorporate graph information



Setup

- Dataset
  Ubuntu IRC benchmark released by Hu et al., 2019

- Baselines
  RNN-based Seq2Seq, Transformer, GPT-2, BERT, GSN and BART

- Metrics
  Automated: BLEU1 to BLEU-4, METEOR and ROUGEL
  Human: relevance, fluency and informativeness

Results

- BERT or BART was selected to initialize the utterance encoder layers of HeterMPC

Models \ Metrics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE$_L$
Seq2Seq (LSTM) (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., 2019b)	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
BERT (Devlin et al., 2019)	10.90	3.85	1.69	0.89	4.18	9.80
HeterMPC$_{BERT}$	**12.61**	**4.55**	**2.25**	**1.41**	**4.79**	**11.20**
HeterMPC$_{BERT}$ w/o. node types	11.76	4.09	1.87	1.12	4.50	10.73
HeterMPC$_{BERT}$ w/o. edge types	12.02	4.27	2.10	1.30	4.74	10.92
HeterMPC$_{BERT}$ w/o. node and edge types	11.60	3.98	1.90	1.18	4.20	10.63
HeterMPC$_{BERT}$ w/o. interlocutor nodes	11.80	3.96	1.75	1.00	4.31	10.53
BART (Lewis et al., 2020)	11.25	4.02	1.78	0.95	4.46	9.90
HeterMPC$_{BART}$	**12.26**	**4.80**	**2.42**	**1.49**	**4.94**	**11.20**
HeterMPC$_{BART}$ w/o. node types	11.22	4.06	1.87	1.04	4.57	10.63
HeterMPC$_{BART}$ w/o. edge types	11.52	4.27	2.05	1.24	4.78	10.90
HeterMPC$_{BART}$ w/o. node and edge types	10.90	3.90	1.79	1.01	4.52	10.79
HeterMPC$_{BART}$ w/o. interlocutor nodes	11.68	4.24	1.91	1.03	4.79	10.65

Table 1: Performance of HeterMPC and ablations on the test set in terms of automated evaluation. Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05).

Models \ Metrics	Score	Kappa
Human	2.81	0.55
GSN (Hu et al., 2019b)	2.00	0.50
BERT (Devlin et al., 2019)	2.19	0.42
BART (Lewis et al., 2020)	2.24	0.44
HeterMPC$_{BERT}$	2.39	0.50
HeterMPC$_{BART}$	2.41	0.45

Table 2: Human evaluation results of HeterMPC and some selected systems on a randomly sampled test set.

Analysis

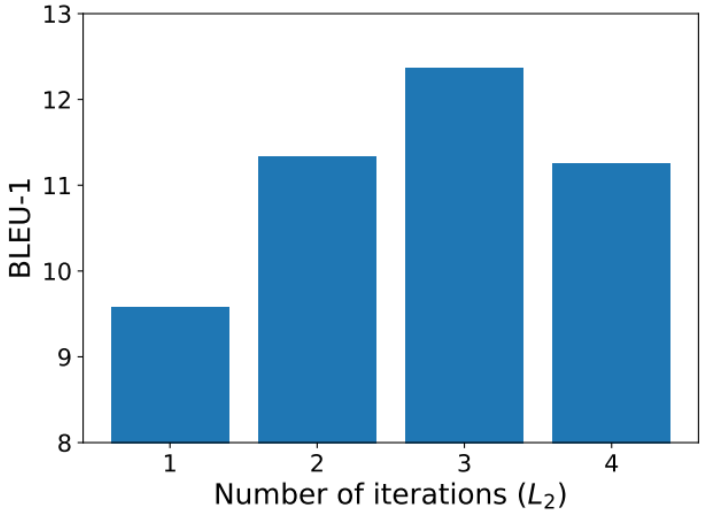


Figure 5: Performance of HeterMPC$_{BERT}$ under different numbers of iterations (L_2) on the test set.

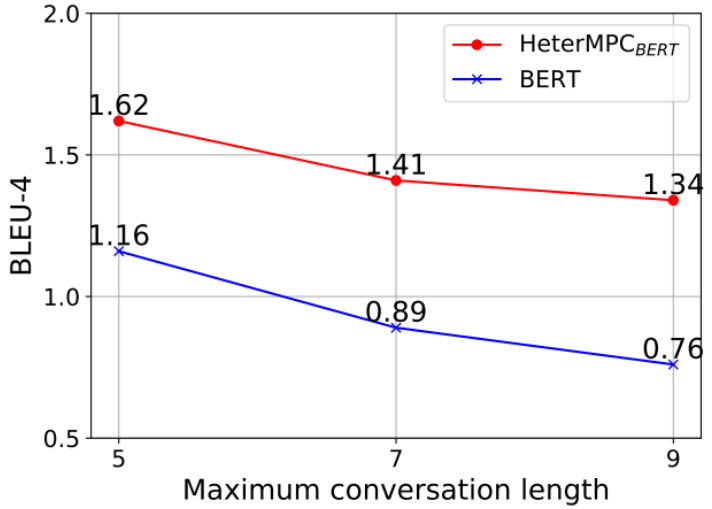


Figure 6: Performance of BERT and HeterMPC$_{BERT}$ on test samples with different session lengths.

The performance of was significantly improved as L2 increased at the beginning. Then, the performance was stable and dropped slightly.

As session lengths increased, the performance dropped less than that of BERT, showing superiority on dealing with longer conversations.

Case Study

- "*gparted*" in the first utterance is two hops away from I.2 (the first utterance is replied by the second utterance which is spoken by I.2), and "*install gparted*" in the third utterance is one hop away from I.2 (this utterance directly addresses I.2).
- These keywords are included in both responses generated by HeterMPC$_{BERT}$ and HeterMPC$_{BART}$, but are missing in those generated by other models.

Speaker	*Utterance*	*Addressee*
	Case 1	
I.1	geev: in gparted now?	-
I.2	there is no such command in my computer	I.1
I.1	open a terminal, type: sudo apt-get install gparted	I.2
I.1	after: gparted aksks for your pw, opens a gui	I.2
I.2	ok if found fail i will contact you	I.1
I.3	does it say "to install type sudo apt-get install gparted"? **(Human)**	I.2
	do you have remote access to remote desktop? **(GSN)**	
	ok, i will try that. **(BERT)**	
	you can use the command "sudoapt-get install gparted" to install the gparted livecd. **(HeterMPC$_{BERT}$)**	
	i'm not sure what you mean by "in gparted" **(BART)**	
	you need to install gparted, it is a gui tool **(HeterMPC$_{BART}$)**	

Outline

- Introduction
- MPC-BERT
- HeterMPC
- **Conclusion**

Conclusion

- We present MPC-BERT
  - a language model with five self-supervised tasks for MPC understanding
  - jointly learn who says what to whom in MPCs
  - outperform previous methods on three downstream tasks

- We present HeterMPC
  - model complicated interactions between utterances and interlocutors in MPCs with a heterogeneous graph
  - two types of graph nodes and six types of edges are designed for better utilizing the structural knowledge of conversations during node updating
  - outperform previous methods on the task of response generation

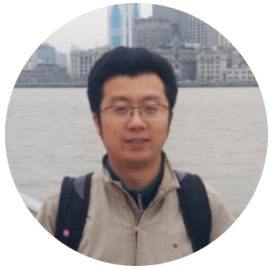
Challenges

- Make up for the performance degradation caused by the lack of addressee labels
- Extend conversations from domain-specific to open-domain
- Unify MPC understanding and generation, and enhance the generalization ability on more downstream tasks, e.g., dialogue disentanglement and discourse parsing

Jia-Chen Gu

Chao-Hong Tan

Chongyang Tao



Zhen-Hua Ling

Huang Hu

Xiubo Geng

Daxin Jiang

MPC-BERT:

HeterMPC:

Thanks! Q&A

Homepage: http://staff.ustc.edu.cn/~gujc/

Contact: gujc@ustc.edu.cn

GitHub: https://github.com/JasonForJoy