
**CCKS 2023 Tutorial**

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

Jia-Chen Gu

University of Science and Technology of China

August 24, 2023

# Presenter

Jia-Chen Gu
Postdoc@USTC

- 2023 - ACL 2023 Best Paper Honorable Mention Award (First-author)
- 2022 - Best Paper Award of ACL 2022 DialDoc Workshop (Second-author)
- 2022 - Outstanding Doctoral Dissertation Nomination Award of CIPS
- 2022 - Presidential Scholarship of Chinese Academy of Sciences (Top 1%)

# History of Conversational AI

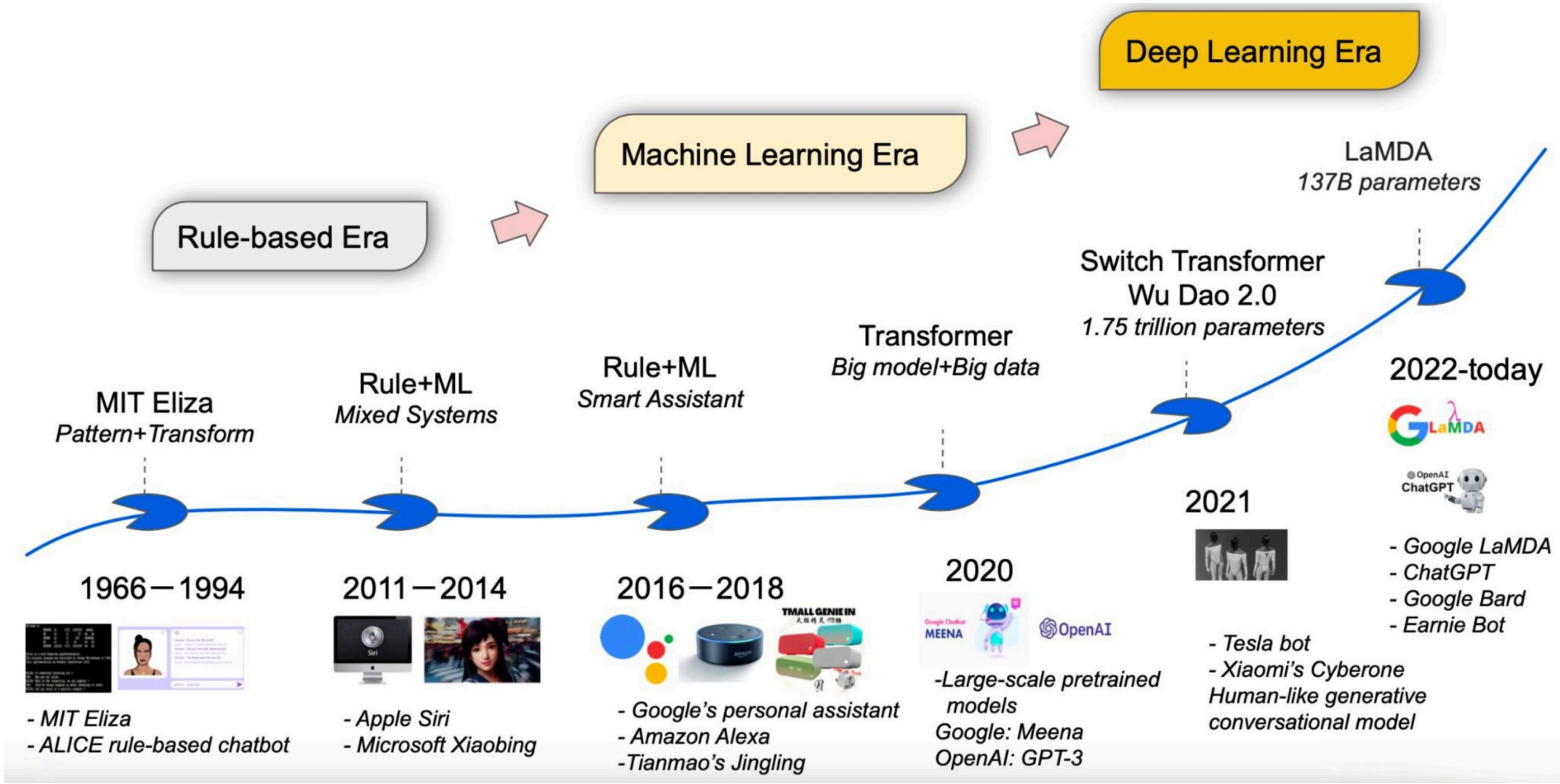
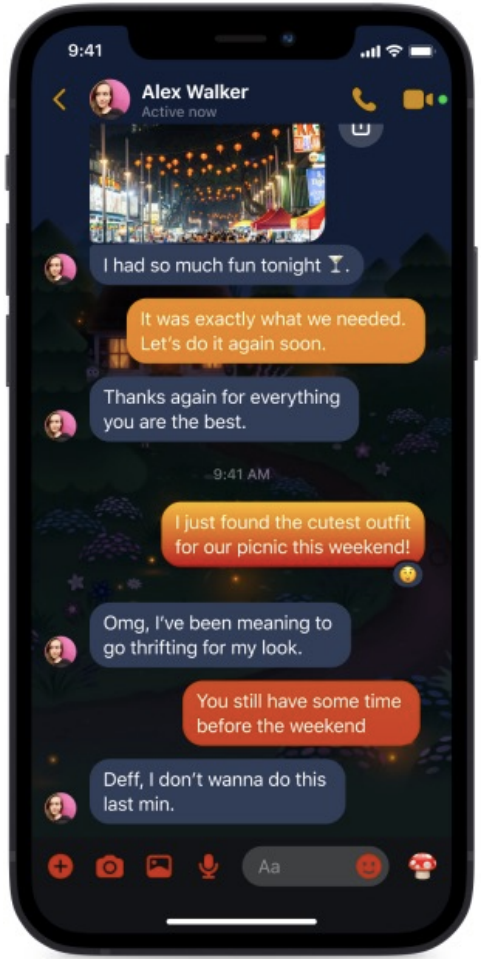



Image: https://dengyang17.github.io/files/ACL2023-Tutorial-ConvAI.pdf

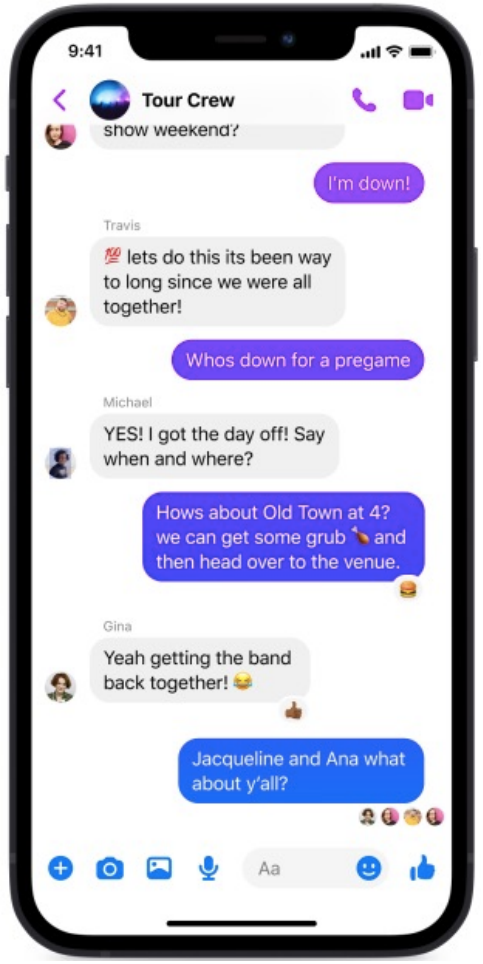
# Two-Party VS. Multi-Party Conversations




One-on-one chat between 2 interlocutors

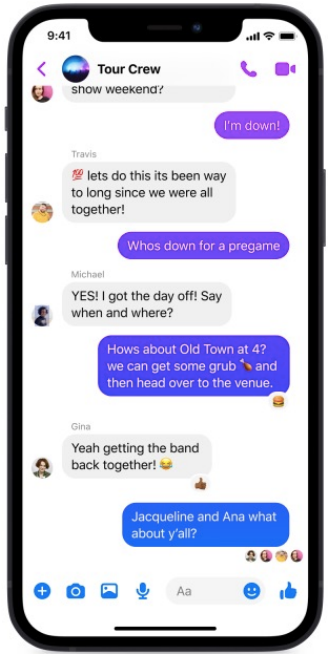
Group chats appear frequently in daily life!

Group chat involving 3+ interlocutors



# Why multi-party conversations (MPC)?

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

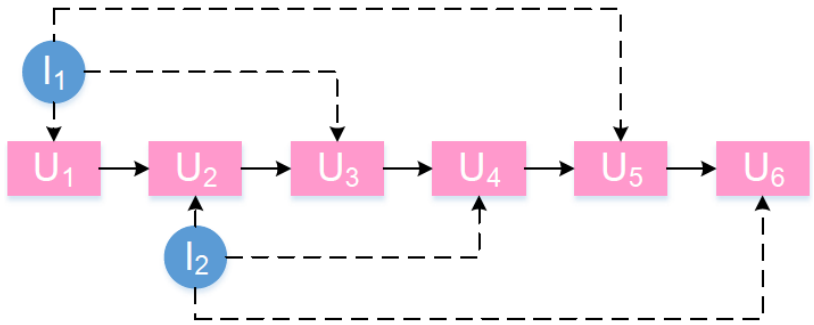


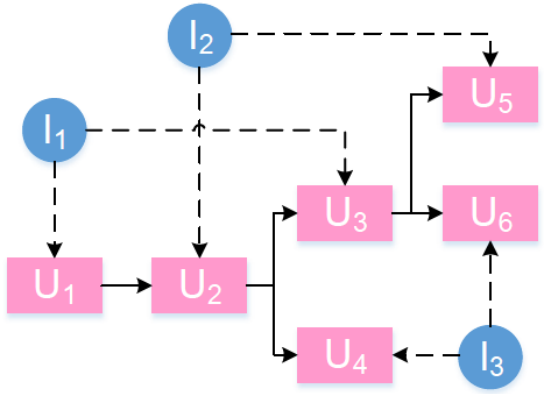

Group Chat

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

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

: Interlocutors

: Utterances

# 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

# Goals of the 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

| Section |
| --- |
| Section 1: Introduction |
| Section 2: Speaker Modeling |
| Section 3: Addressee Modeling |
| Section 4: Response Modeling |
| Section 5: Challenges & Opportunities |
| Q & A Session |

# Section 2:
# Speaker Modeling

# WHO speaks

| 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? ] | | |

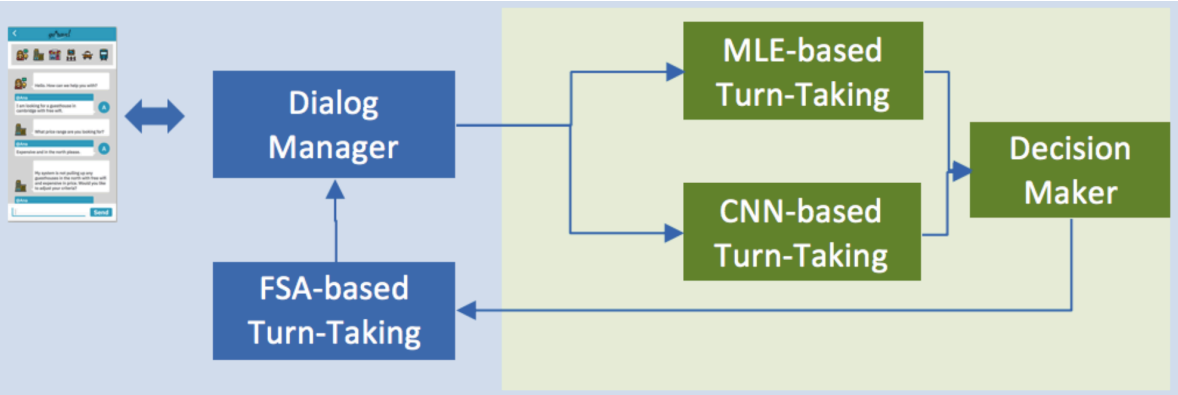
Should the agent take
the floor to speak or not?

# 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

# 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




$$s_{t+1} = \ell(x(t+1)) | \ell \in L$$

$$\ell(x(t+1)) = \begin{cases} \ell(x_1(t+1)) & \text{if } C_1 >= k_1 \\ \ell(x_2(t+1)) & \text{if } C_2 < k_1 \\ & \text{and } \quad C_2 >= k_2 \\ travel\_bot & otherwise \end{cases}$$



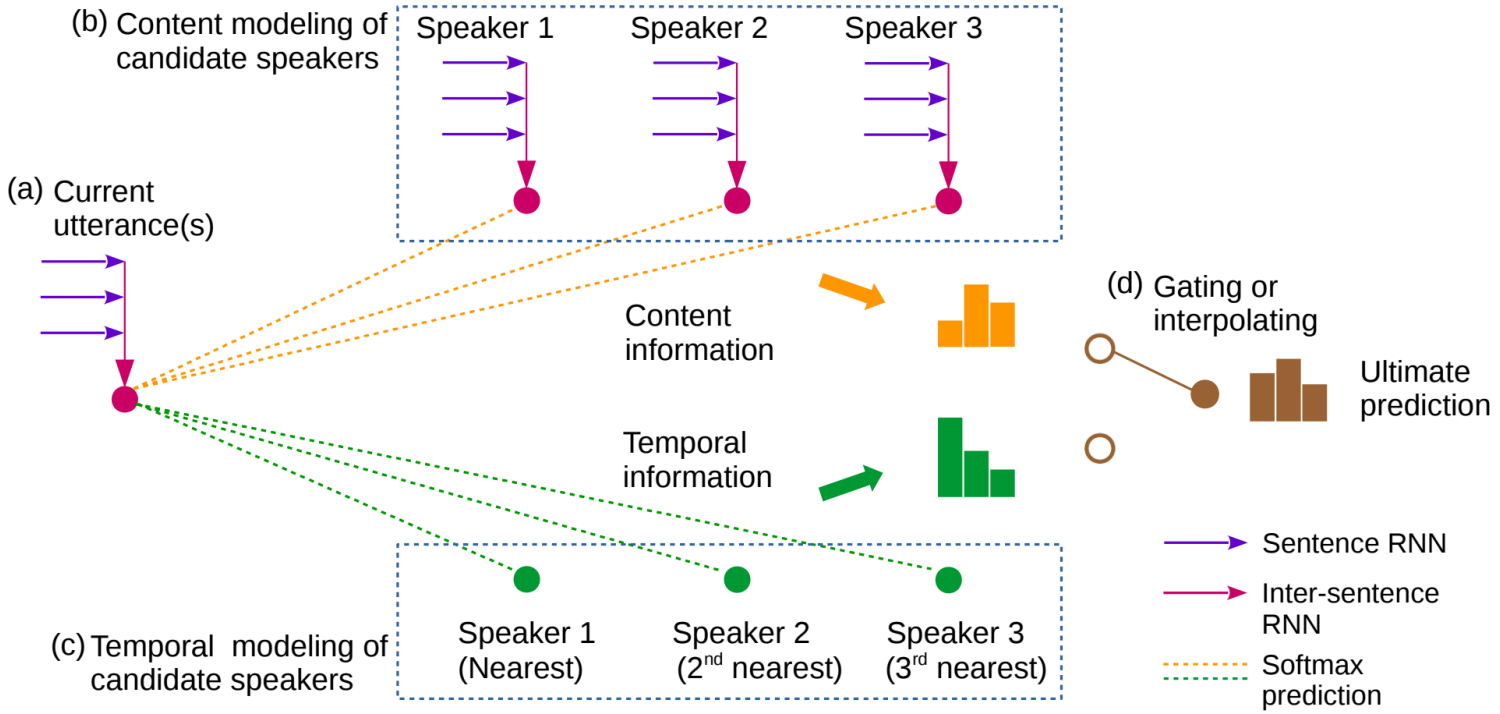
Maira Gatti de Bayser, et al. *A Hybrid Solution to Learn Turn-Taking in Multi-Party Service-based Chat Groups*. 2020.

# 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.

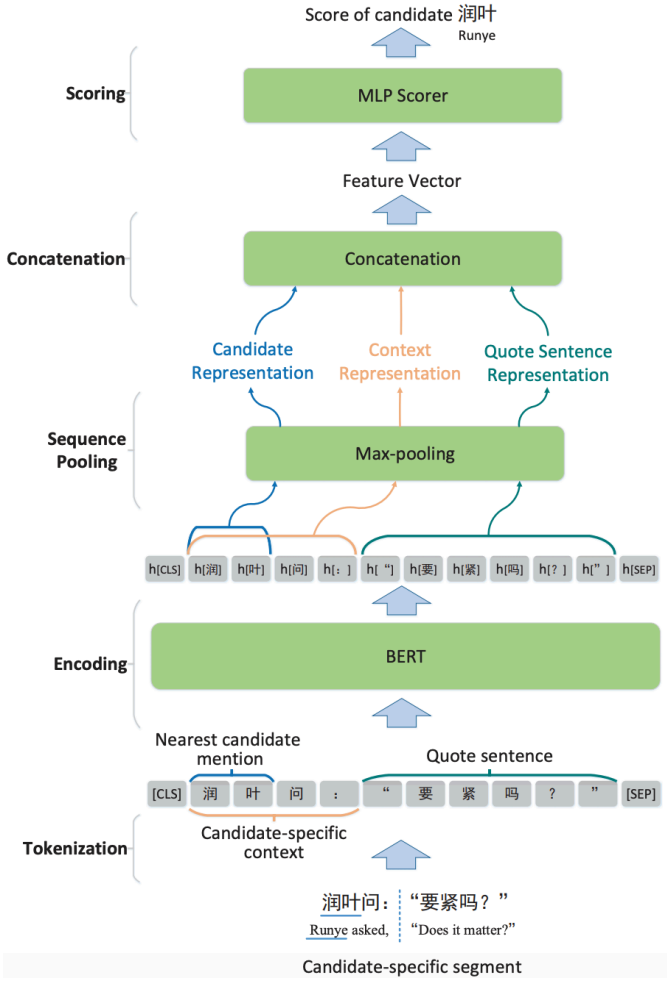
Zhao Meng, et al. *Towards neural speaker modeling in multi-party conversation: The task, dataset, and models.* LREC 2018.

# 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





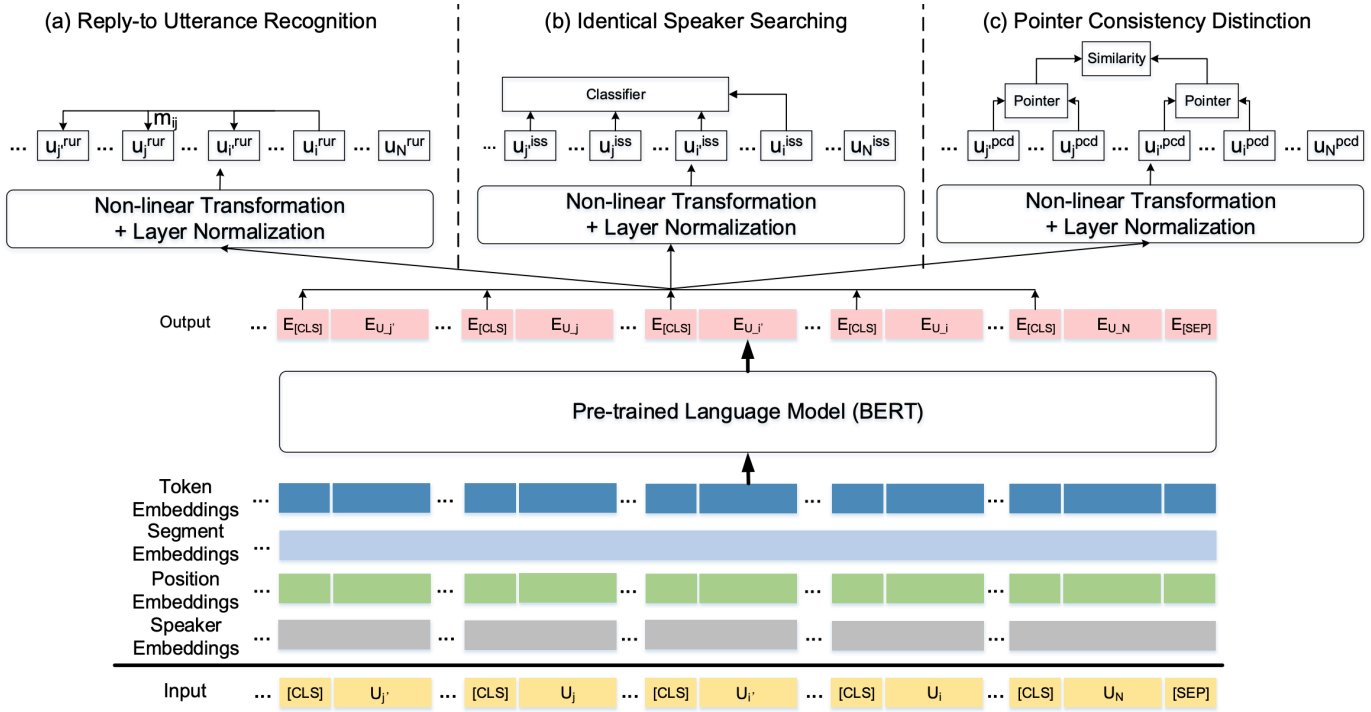

Yue Chen, et al. *A neural-network-based approach to identifying speakers in novels*. INTERSPEECH 2021.

# 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

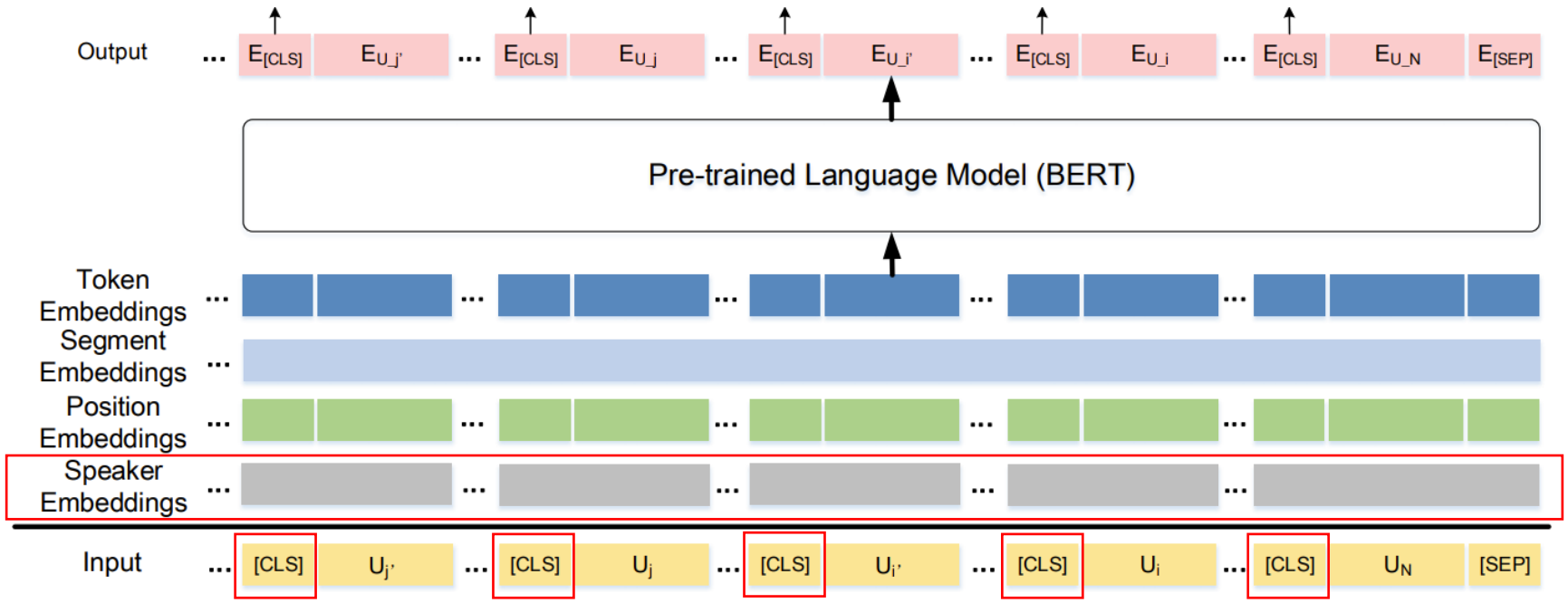





Jia-Chen Gu, et al. *MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding*. ACL 2021.

# 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

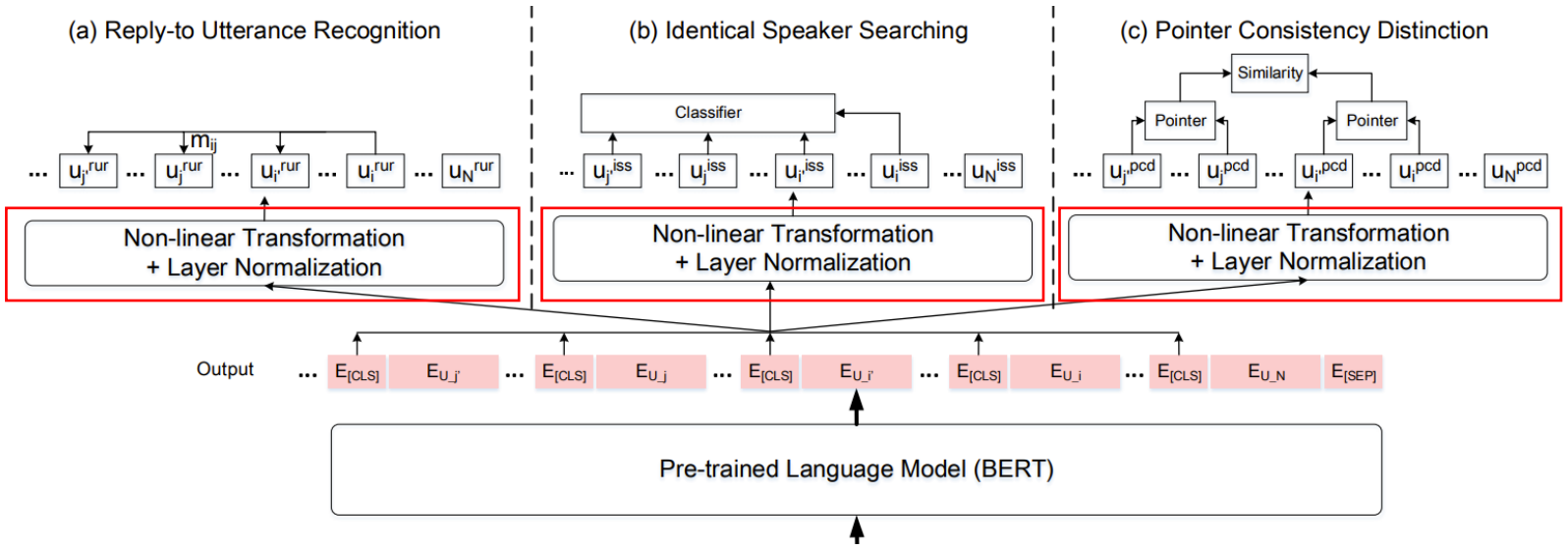





Jia-Chen Gu, et al. *Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots.* CIKM 2020.

# MPC-BERT: interlocutor structure modeling

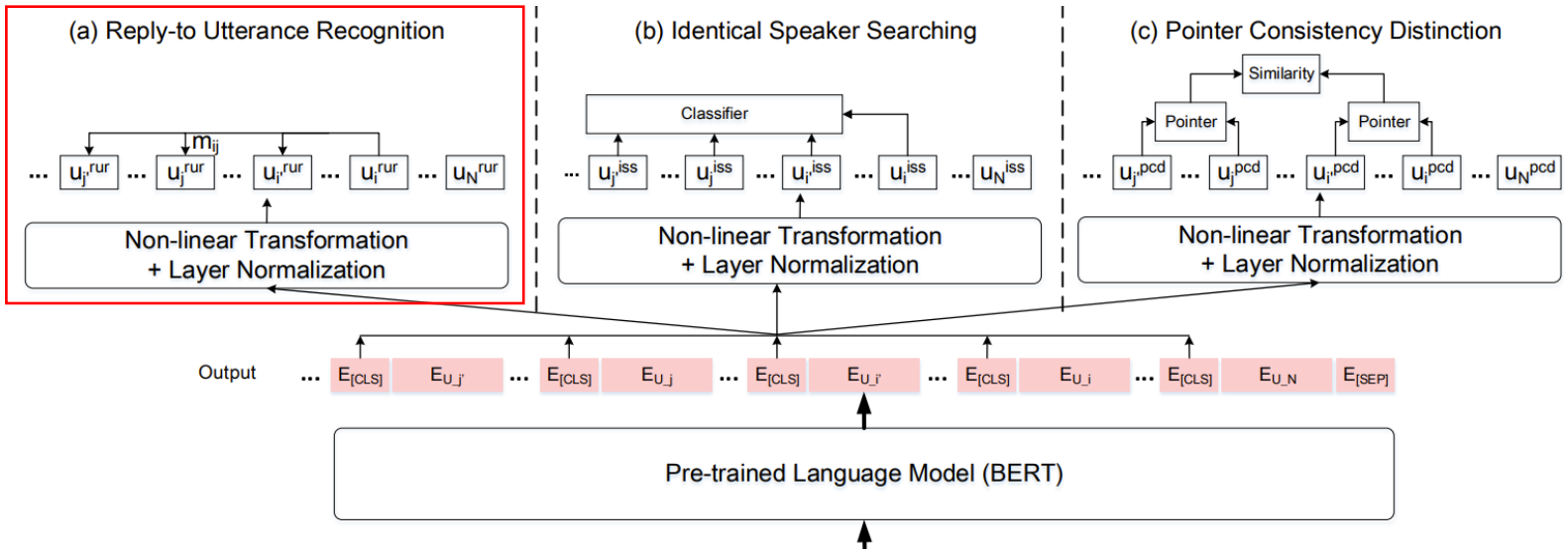
- Extract the representations for each [CLS] token representing utterances
- Task-dependent non-linear transformations are placed on top of BERT for three self-supervised tasks
- Encoding the input data only once is computation-efficient




Utterance semantics modeling part will be covered in Section 4!

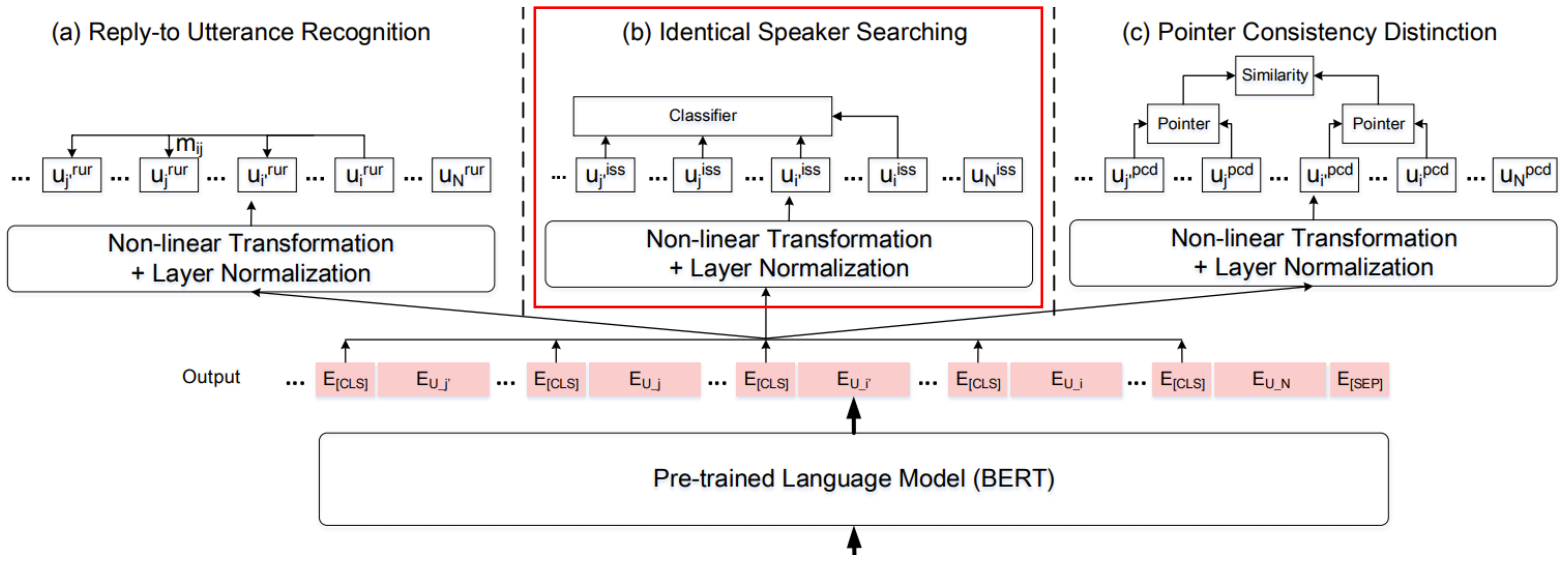
# Reply-to Utterance Recognition

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



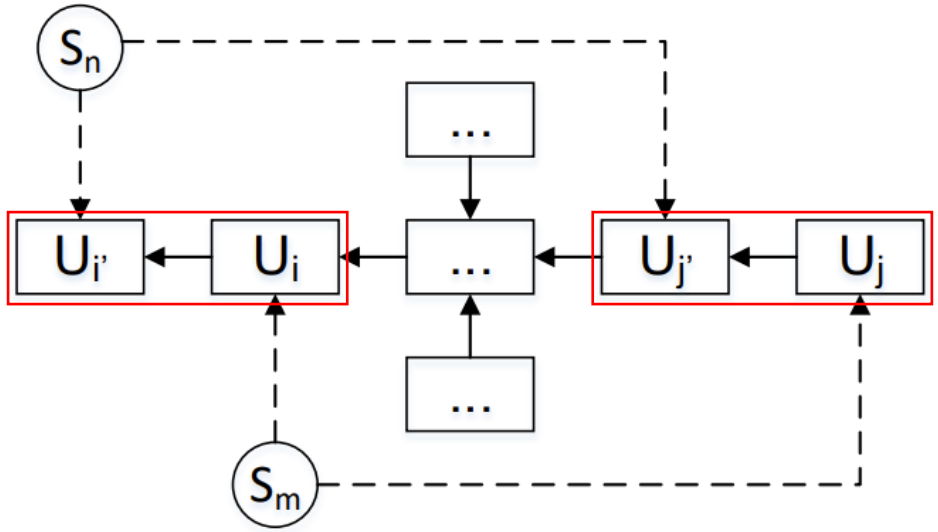
# Identical Speaker Searching

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



# Pointer Consistency Distinction

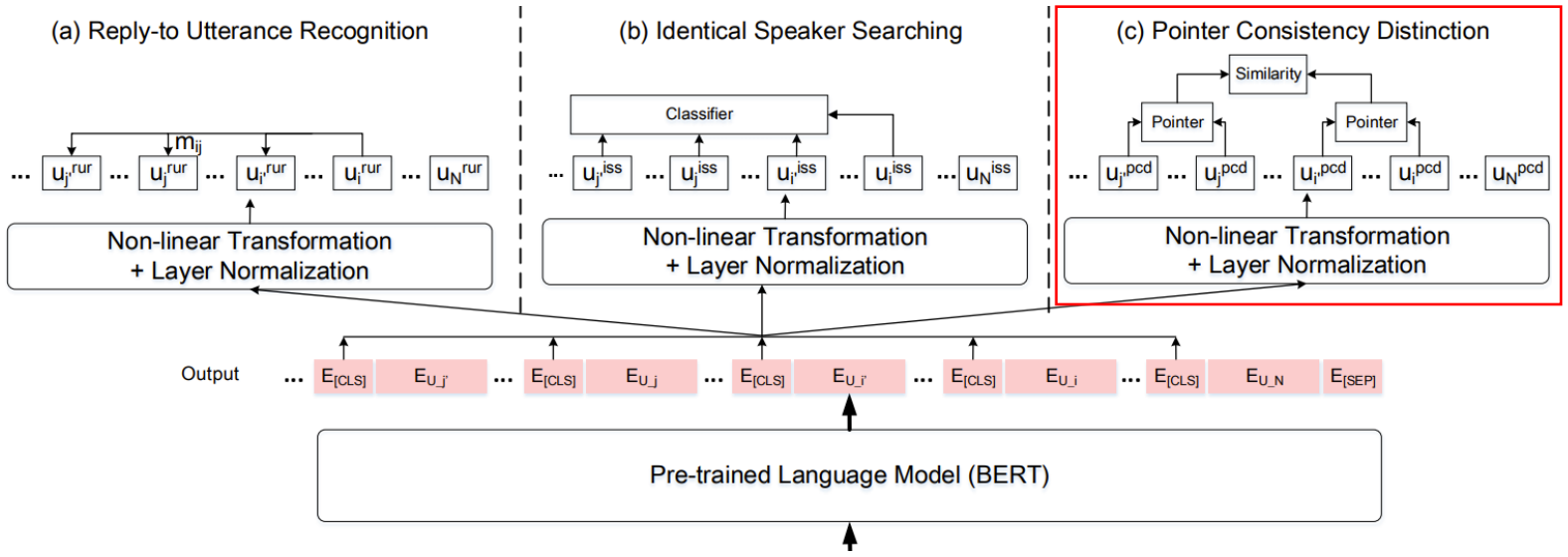
- **Definition**: a speaker-to-addressee pointer is expressed as a pair of utterances representing the "reply-to" relationship
- **Assumption**: the representations of two pointers directing from the same speaker to the same addressee should be consistent



# Pointer Consistency Distinction

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

  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 3:
# Addressee Modeling

# Address WHOM

| 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?

# 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

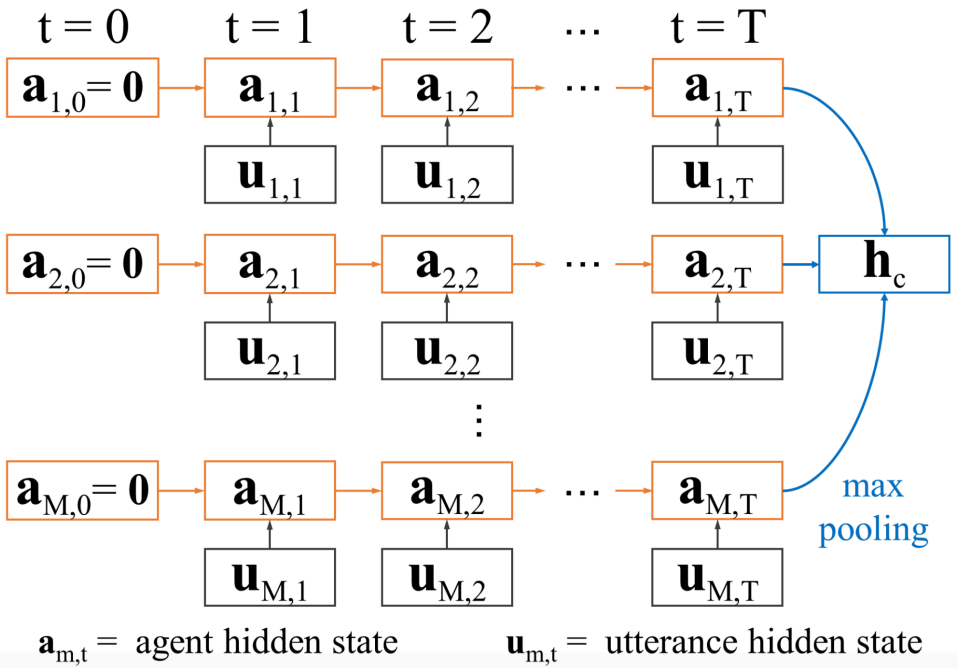
| Speaker | Utterance | Addressee |
|---|---|---|
| User 1 | ”Good point, tmux is the thing I miss.” | – |
| User 1 | ”Cool thanks for ur help.” @User 4 | User 4 |
| User 2 | ”Ahha, you r using something like cpanel.” | – |
| User 3 | ”Yeah 1.4.0 exactly.” @User 2 | User 2 |
| User 4 | ”my pleasure :)”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

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




The states of the agents that

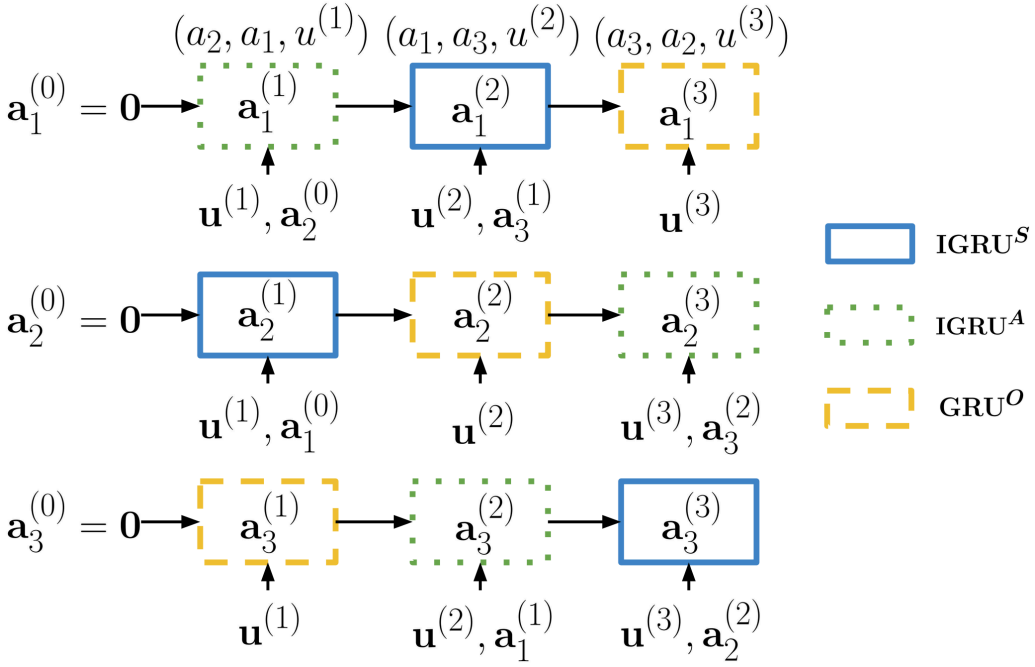
- are speaking at the time are updated by consuming the utterance vector
- are not speaking at the time are updated by zero vectors



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

# 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

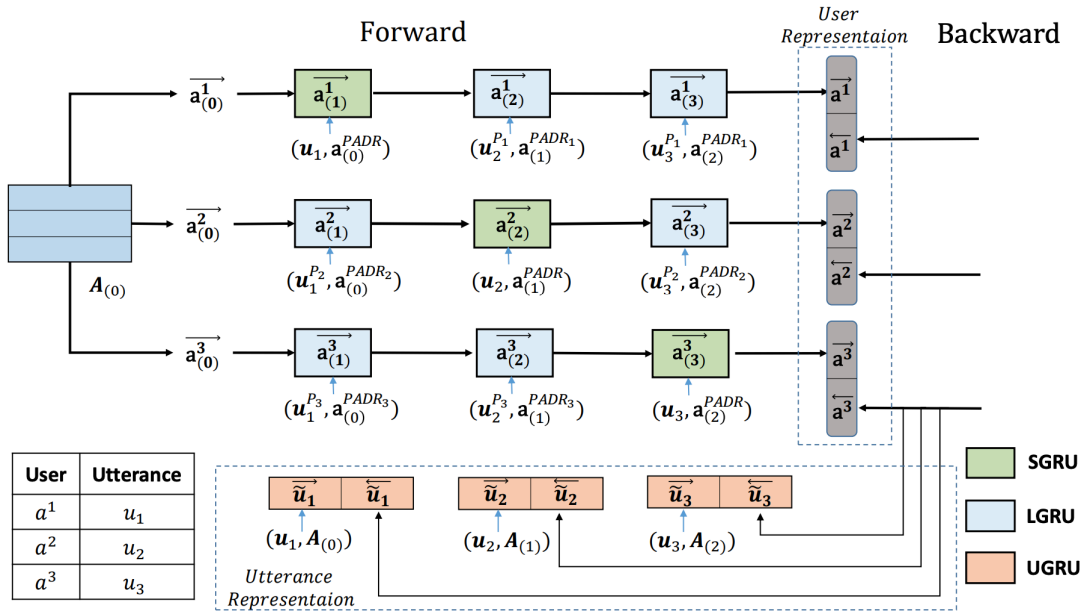
- IGRU$^S$ for sender
- IGRU$^A$ for addressee
- GRU$^O$ for observer



Rui Zhang, et al. *Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs.* AAAI 2018.

# Who-to-Whom

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




| User | Utterance |
|---|---|
| $a^1$ | $u_1$ |
| $a^2$ | $u_2$ |
| $a^3$ | $u_3$ |

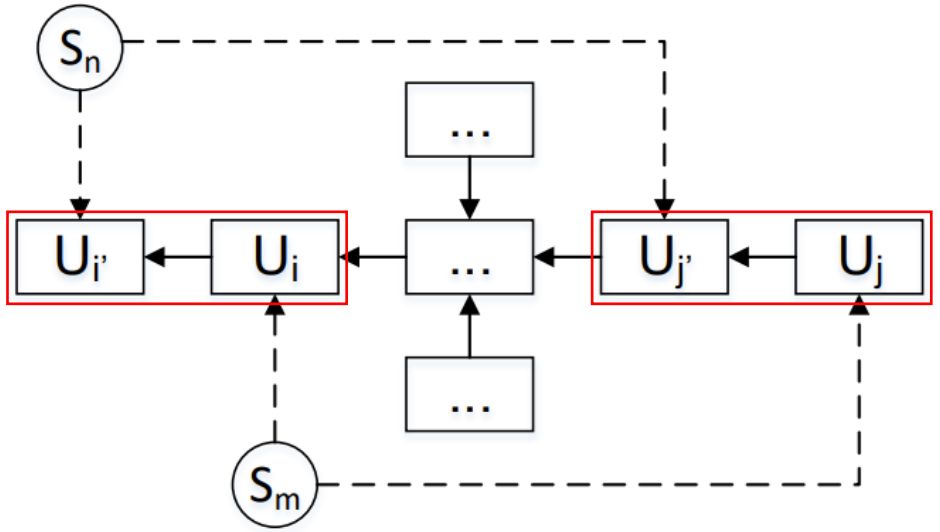
- Track users' states with utterance embeddings (DRNN, SI-RNN), i.e., utterance -> user
- Fuse users' states into the utterance embeddings, i.e., user -> utterance



Ran Le, et al. *Who Is Speaking to Whom? Learning to Identify Utterance Addressee in Multi-Party Conversations.* EMNLP 2019.

# 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



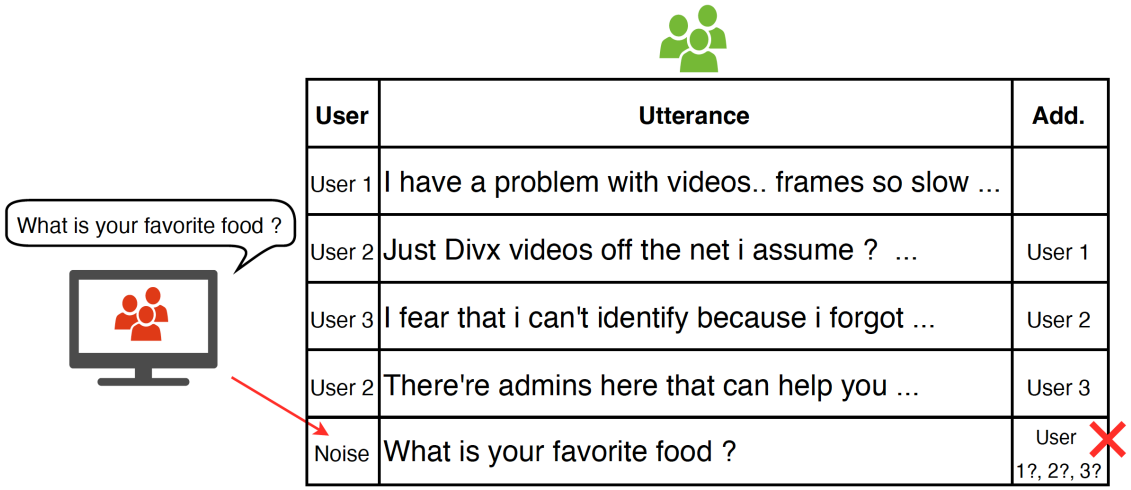

Discussed in Section 2!



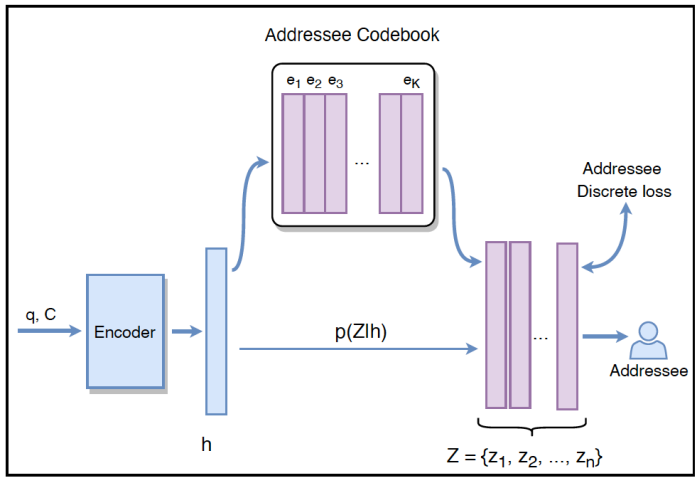
Jia-Chen Gu, et al. *MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding*. ACL 2021.

# 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? |



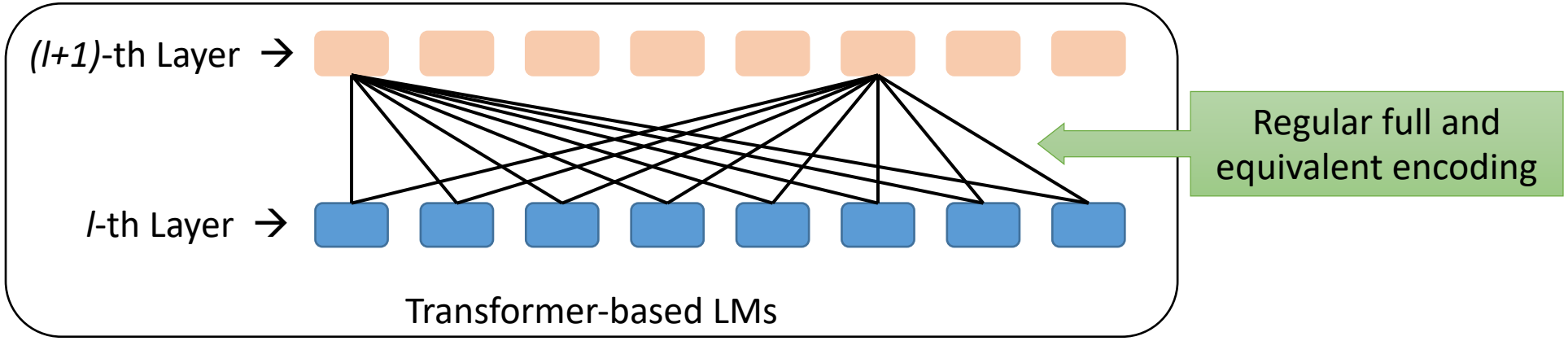

Discretize addressees into a codebook with VQ-VAE to solve the issue of unknown number of addressees in noisy environment



Pengcheng Zhu, et al. *Robust Learning for Multi-party Addressee Recognition with Discrete Addressee Codebook*. ACL 2023.

# GIFT

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



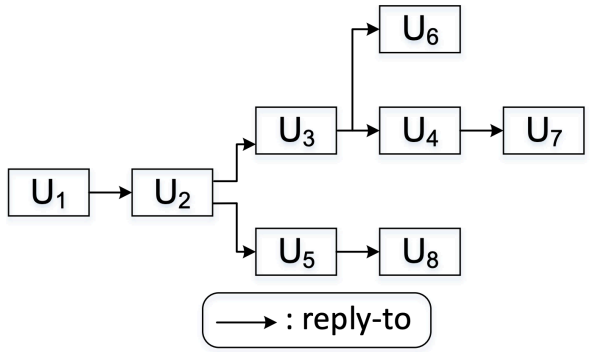

- **Methodology**: to distinguish different utterance relationships for modeling the inherent MPC graph structure via graph-induced fine-tuning



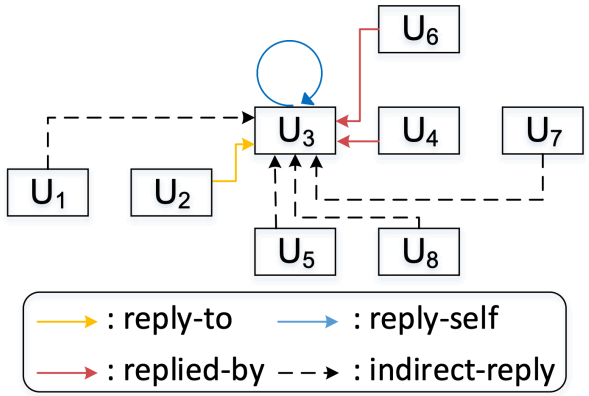
Jia-Chen Gu, et al. *GIFT: Graph-Induced Fine-Tuning for Multi-Party Conversation Understanding*. ACL 2023.

# GIFT Graph Topology

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




(a) A Graphical Information Flow of an MPC




(b) Reply Relationships in a Graph Structure for $U_3$

* Rectangles ( U ) denote utterances, and solid lines (⟶) 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

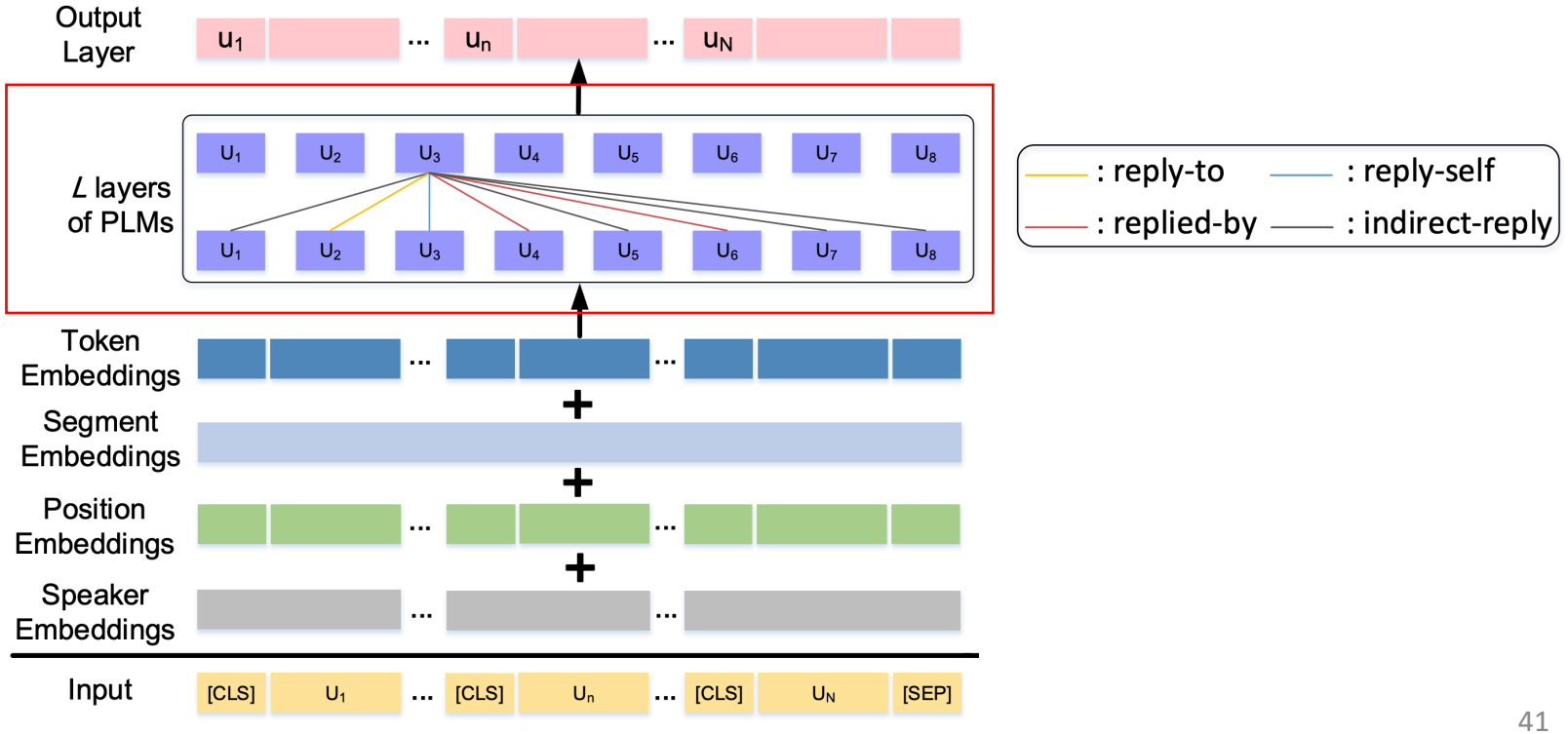
$$\text{Atten}(q, k, v) = \text{softmax}(\phi(e_{q,v})\frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}})\mathbf{v}$$

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

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

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

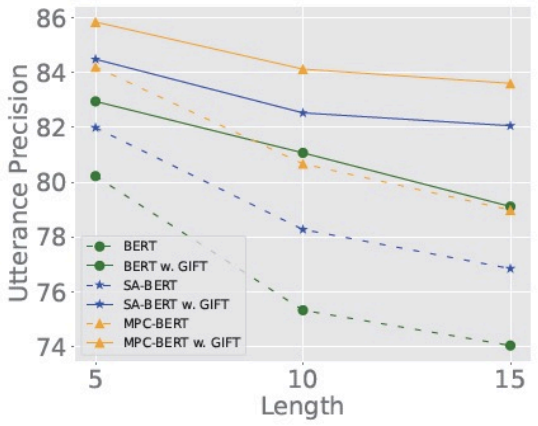
| | Hu et al. (2019) | Ouchi and Tsuboi (2016) | | |
|---|---|---|---|---|
| | | Len-5 | Len-10 | Len-15 |
| BERT (Devlin et al., 2019) | 71.81 | 62.24 | 53.17 | 51.58 |
| SA-BERT (Gu et al., 2020) | 75.88 | 64.96 | 57.62 | 54.28 |
| MPC-BERT (Gu et al., 2021) | 83.54 | 67.56 | 61.00 | 58.52 |
| BERT 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

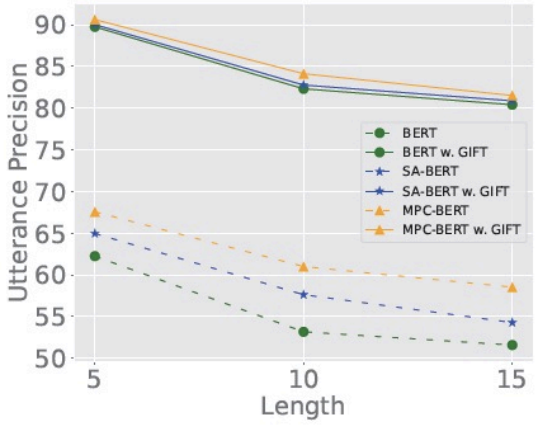
Surprisingly effective for speaker identification!

# Performance Change at Different Lengths

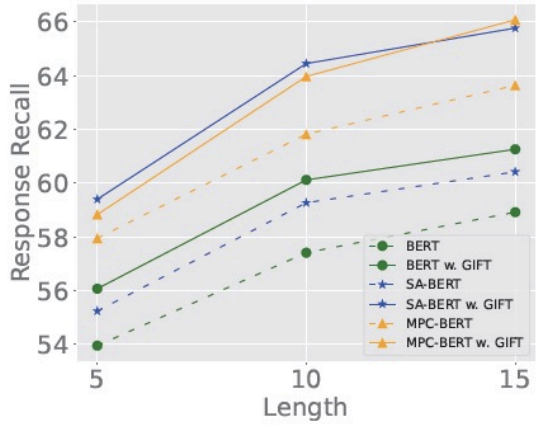
As the session length increased, the performance of models with GIFT dropped more slightly on addressee recognition and speaker identification, and enlarged more on response selection, than the models without GIFT in most 14 out of 18 cases




(a) Addressee Recognition




(b) Speaker Identification



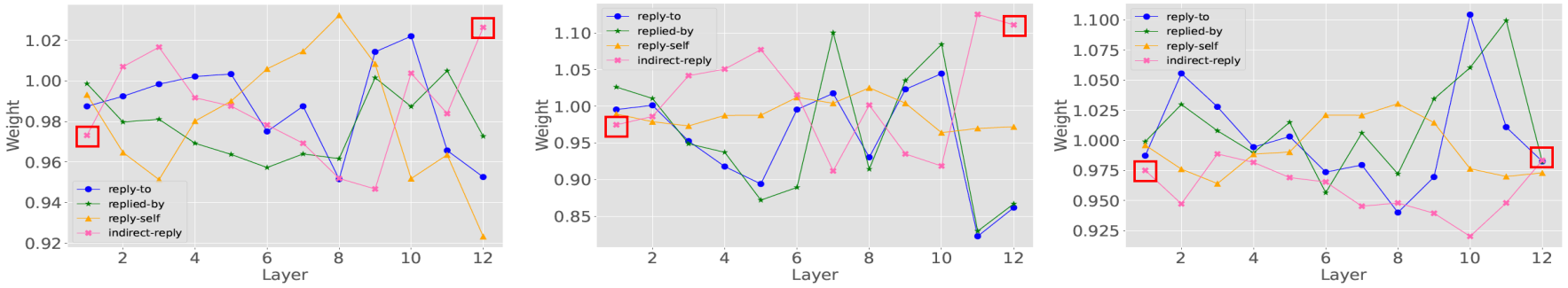

(c) Response Selection

| | 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).


# We will be back!

# Section 3:
# 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

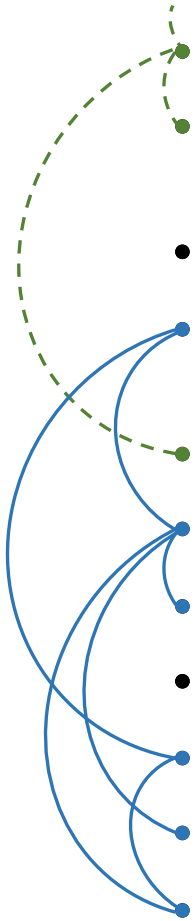
- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...@...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?.. this digital camera age kills me 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



Jonathan K. Kummerfeld, et al. *A Large-Scale Corpus for Conversation Disentanglement*. ACL 2019.

# Dialogue Disentanglement



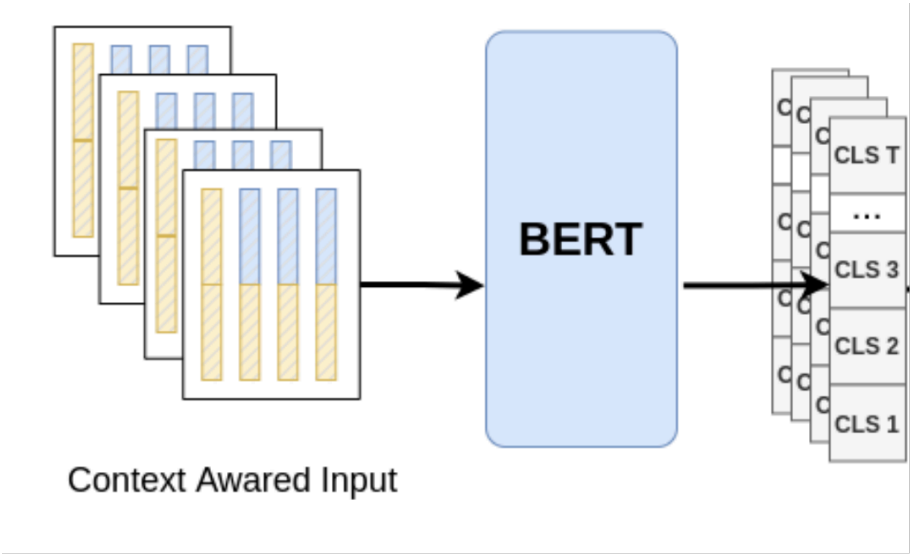
- [03:05] hehe yes. does Kubuntu have 'KPackage'?
- === delire found that to be an excellent interface to the apt suite in another distribution.
- === E-bola [...@...] has joined #ubuntu
- [03:06] does anyone know a consoleprog that scales jpegs fast and efficient?.. this digital camera age kills me 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



Jonathan K. Kummerfeld, et al. *A Large-Scale Corpus for Conversation Disentanglement*. ACL 2019.

# DialBERT

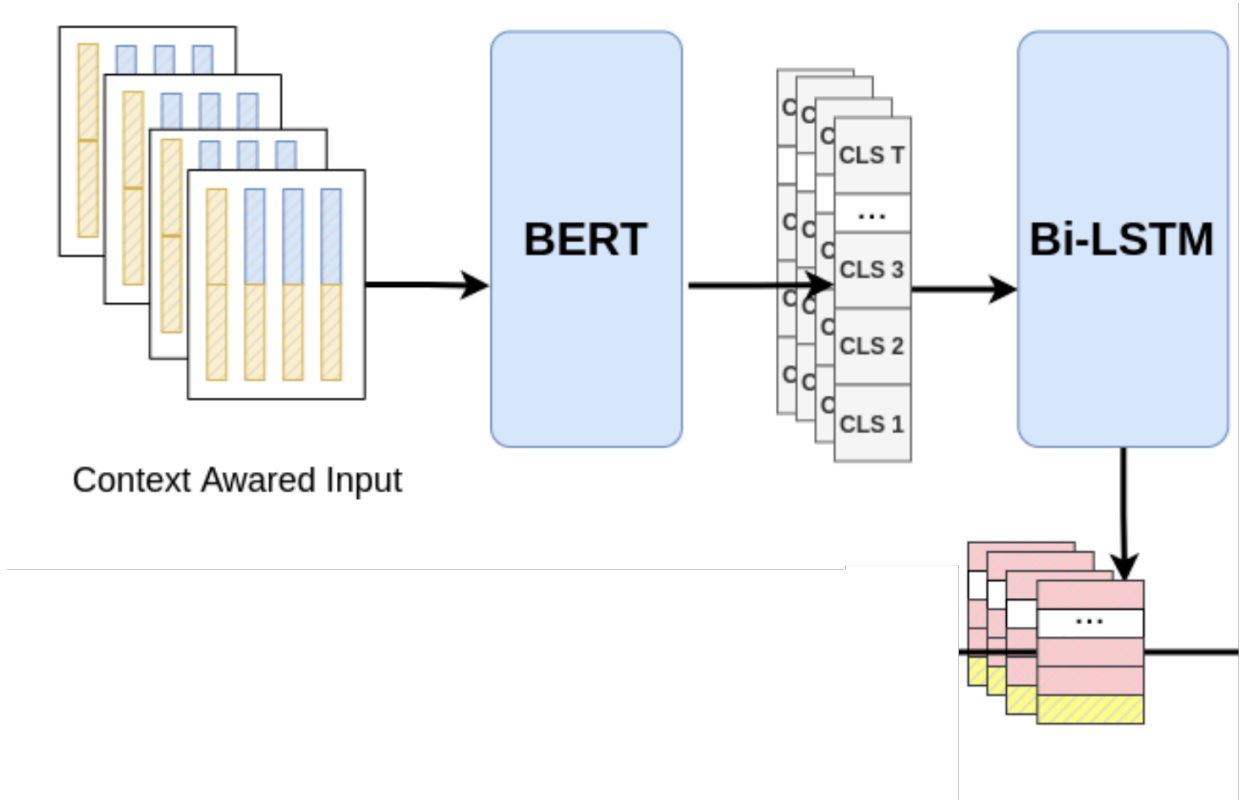



Capture local semantics by concatenating target with each of context utterance

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

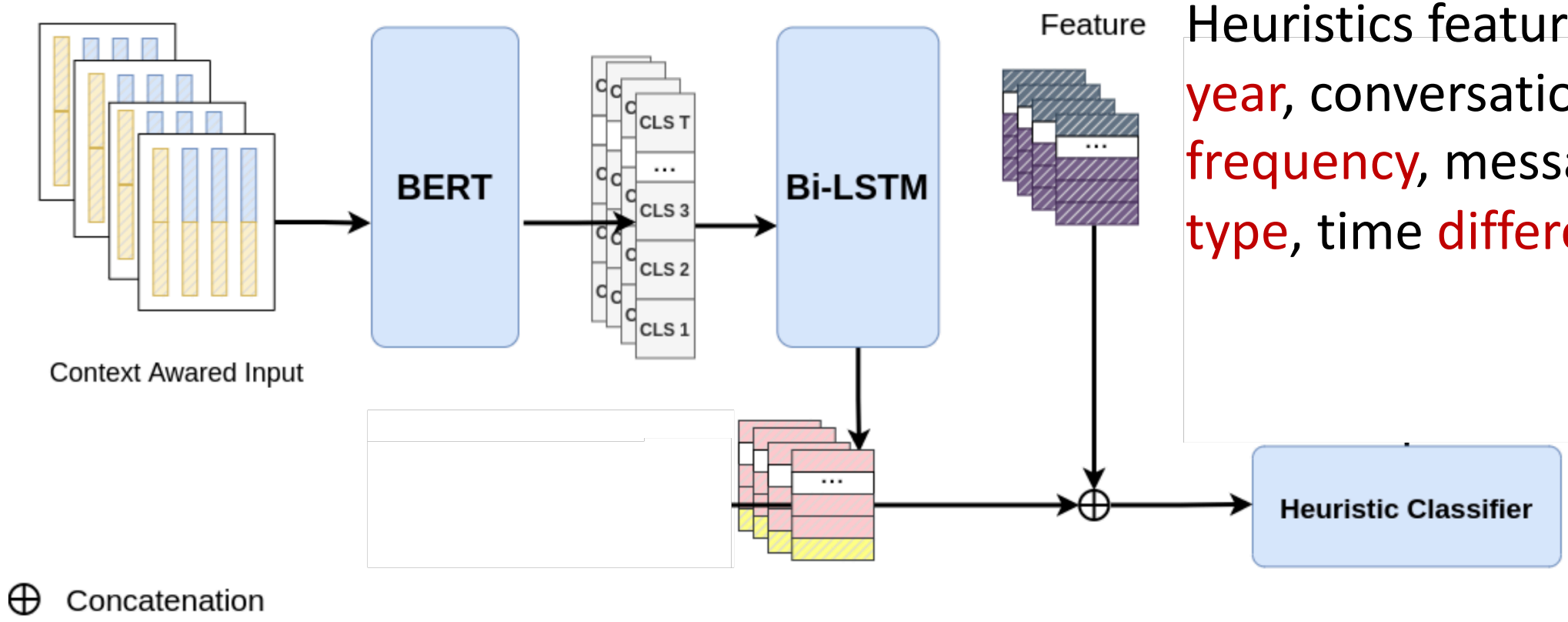



Capture global semantics across different message pairs to enhance context

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



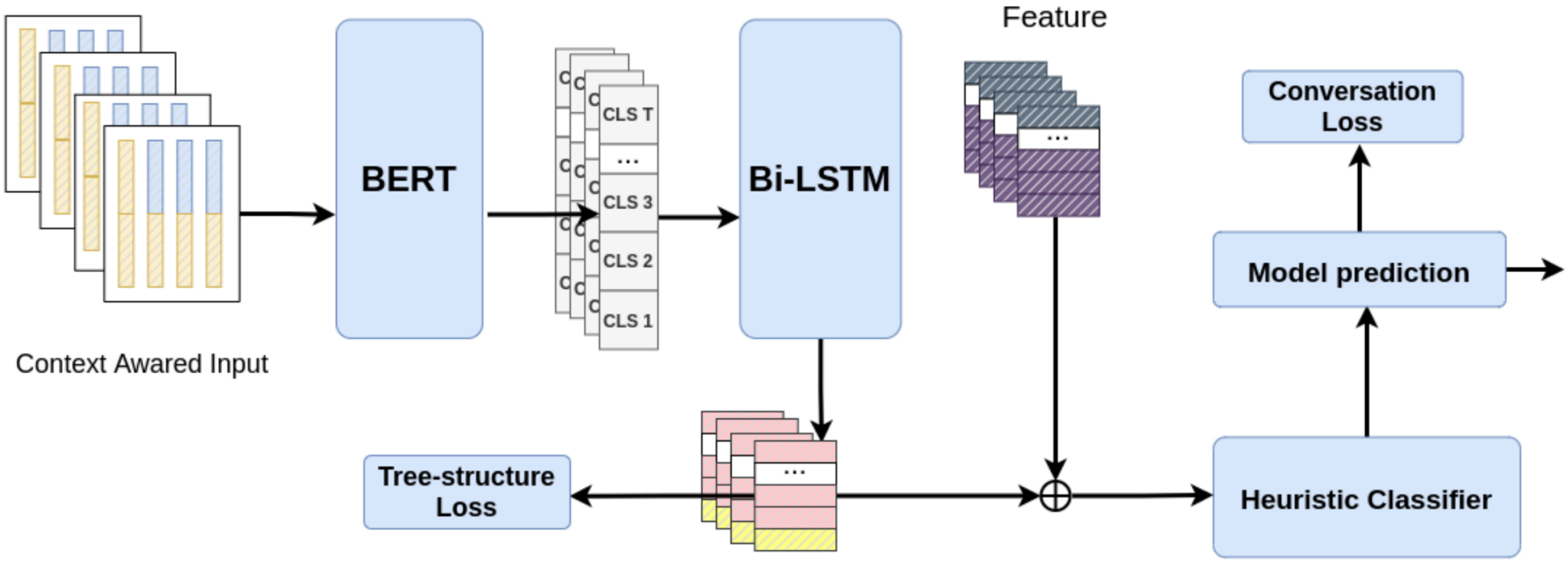

Heuristics features, e.g., year, conversation frequency, message type, time difference

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




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.

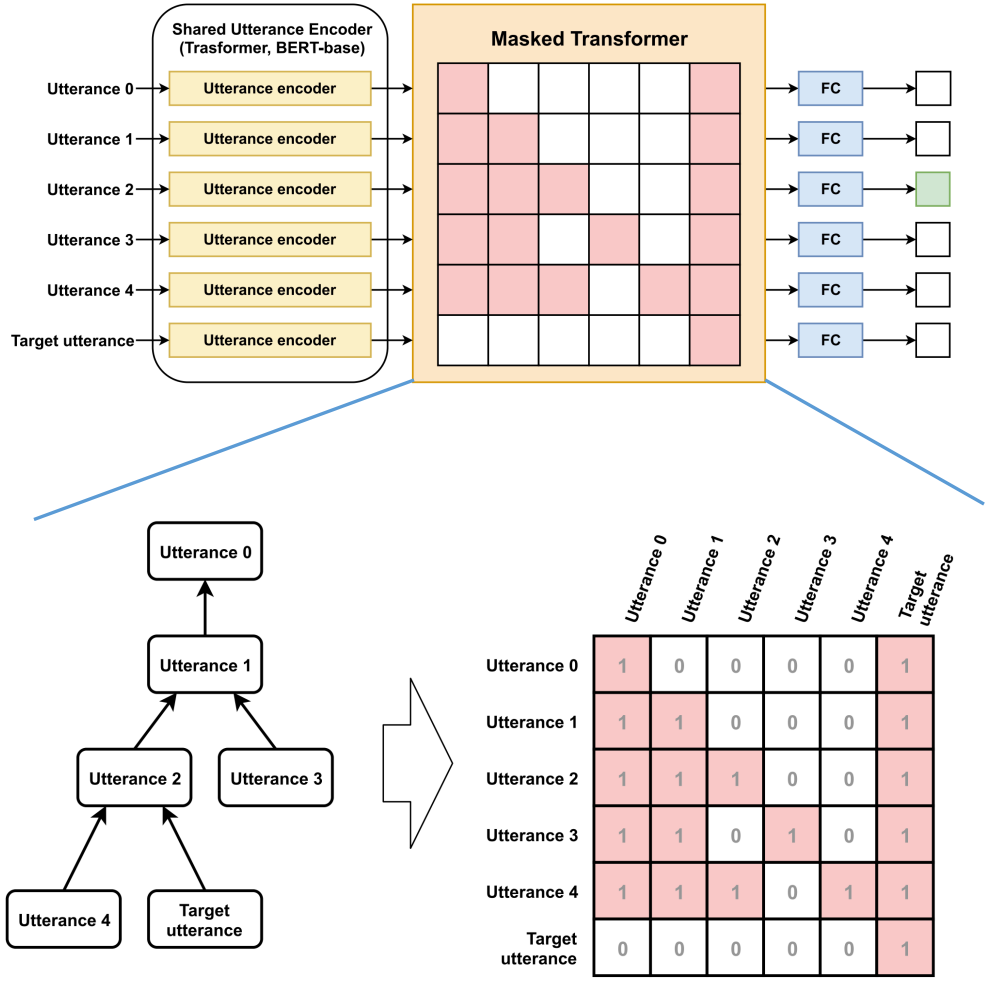
Tianda Li, et al. *Conversation- and Tree-Structure Losses for Dialogue Disentanglement*. DialDoc 2022.

# Masked Hierarchical Transformer

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

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






Henghui Zhu, et al. *Who did They Respond to? Conversation Structure Modeling using Masked Hierarchical Transformer*. AAAI 2020.

# Section 4:
# Response Modeling

# Say WHAT

| 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? ] |

# 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)

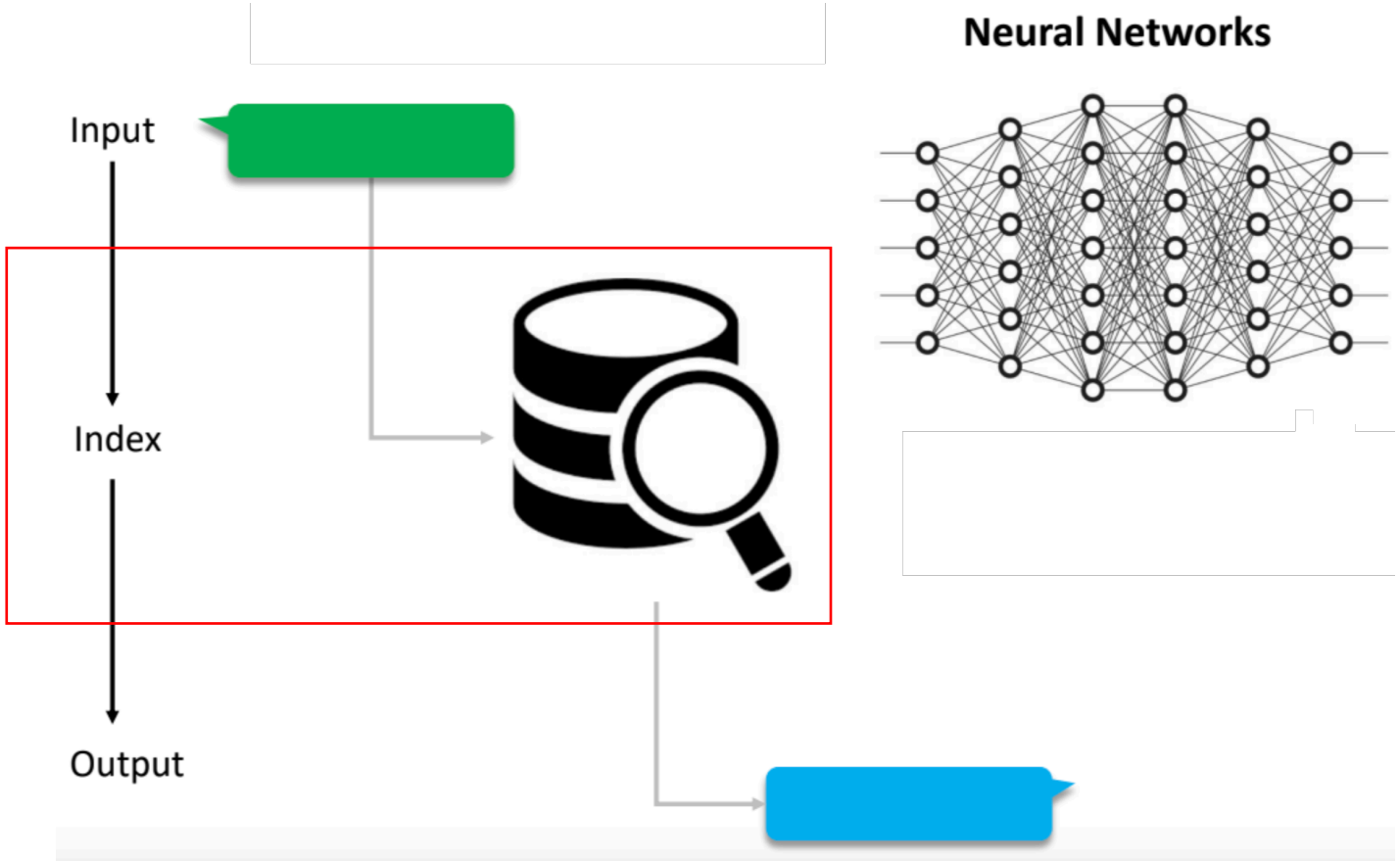
# Response Selection

Not only:

- Semantics
- Consistency
- Interactiveness

But also:

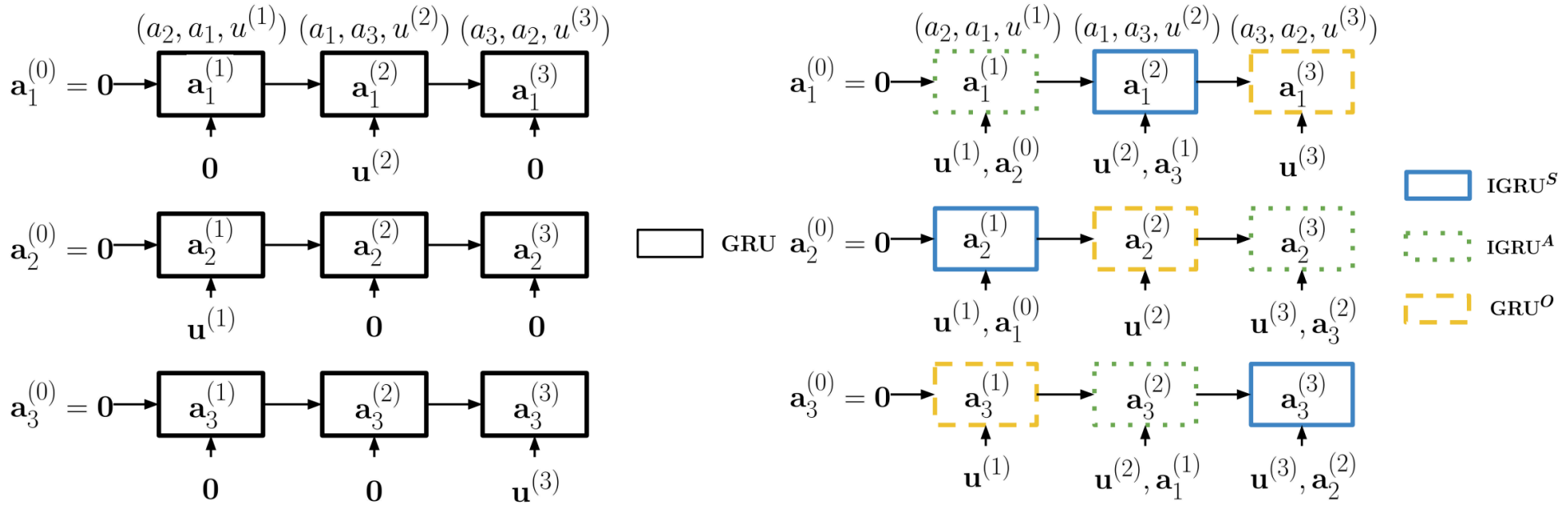
- Conversation structure
- Topic transition



# 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.

Rui Zhang, et al. *Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs.* AAAI 2018.

# Topic-BERT

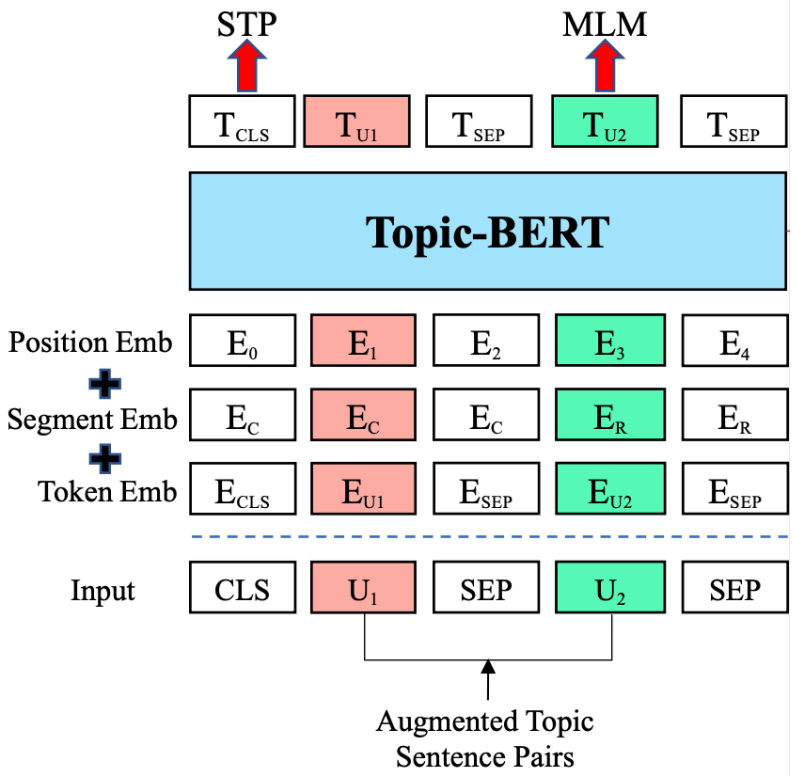
Frame response selection as dynamic topic tracking

→ remain the same topic as going from context to response



Weishi Wang, et al. *Response Selection for Multi-Party Conversations with Dynamic Topic Tracking*. EMNLP 2020.

# Topic-BERT




Pretrain with

a) Same Topic Prediction (STP): if a pair of utterance in a single-threaded conversation

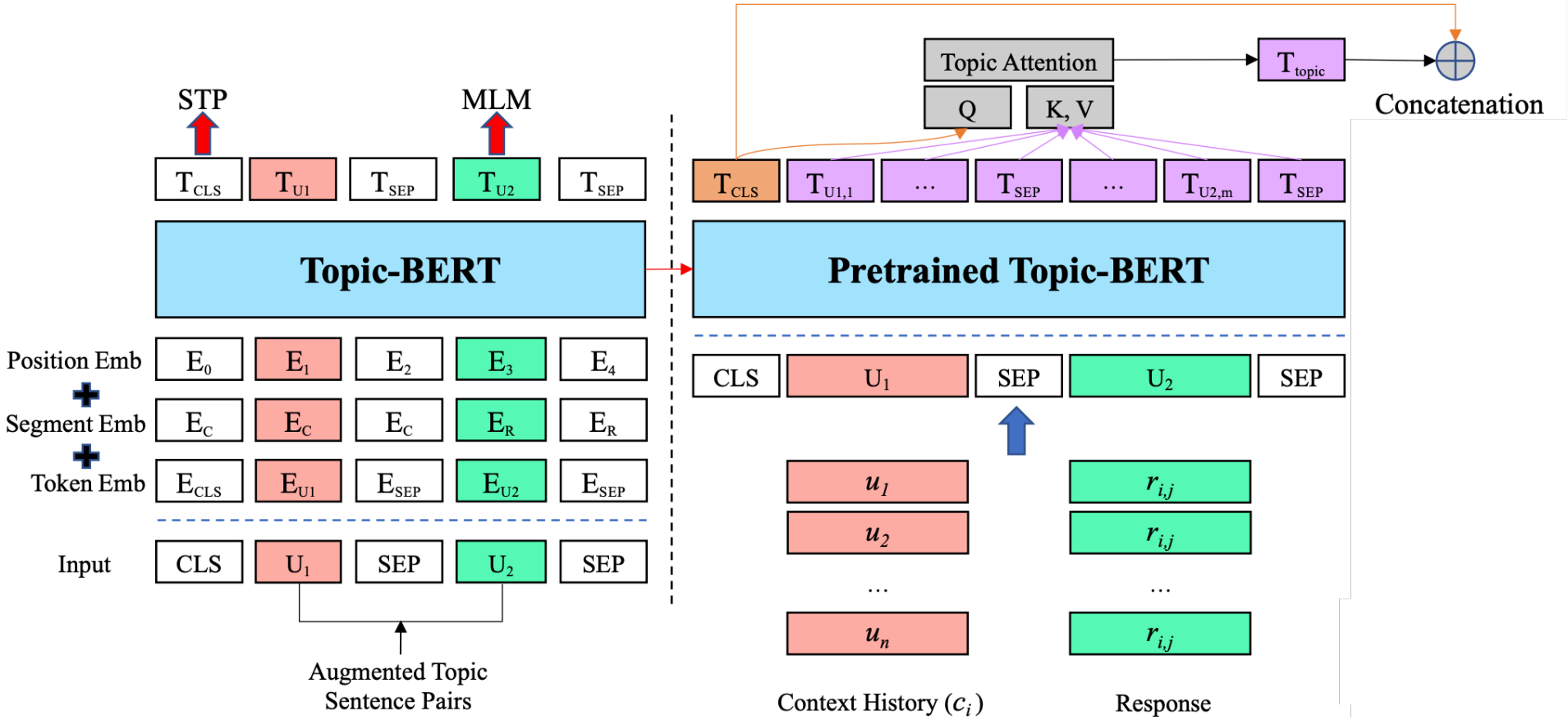
b) Masked Language Modeling (MLM)



Weishi Wang, et al. *Response Selection for Multi-Party Conversations with Dynamic Topic Tracking*. EMNLP 2020.

# Topic-BERT

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

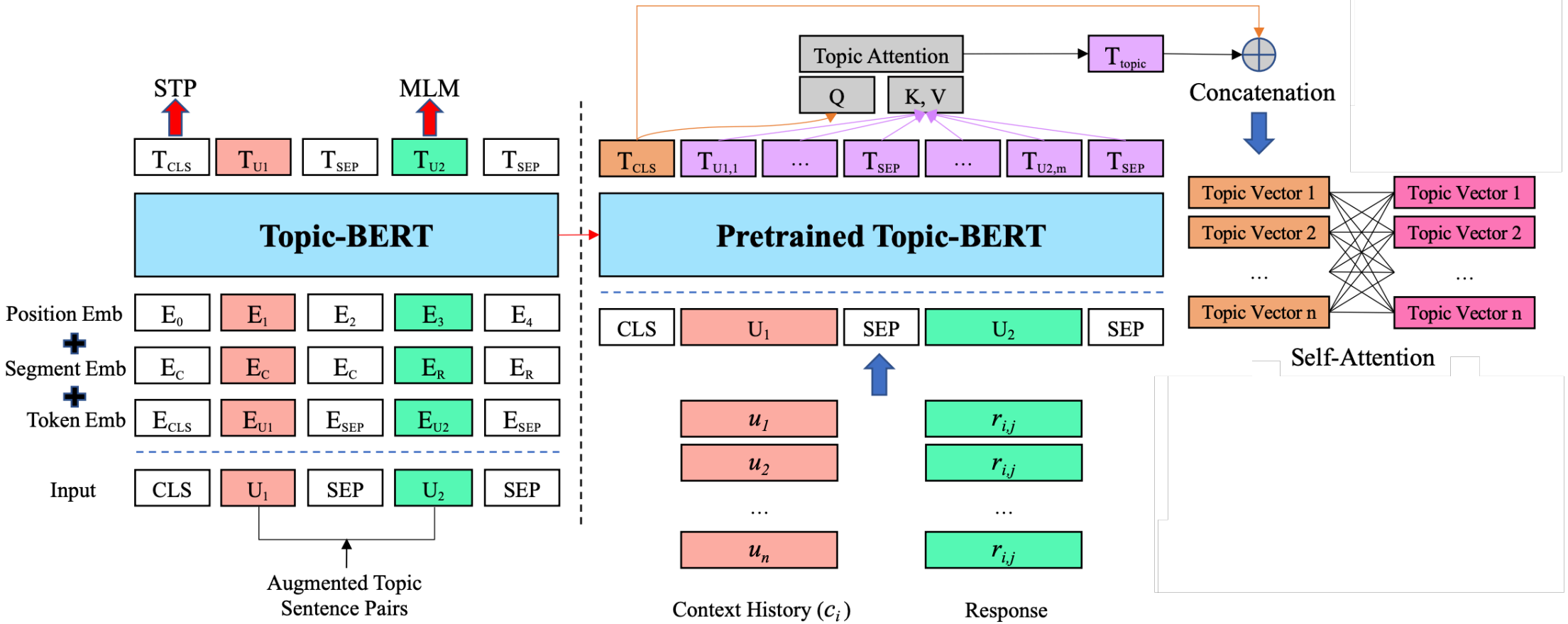





Weishi Wang, et al. *Response Selection for Multi-Party Conversations with Dynamic Topic Tracking*. EMNLP 2020.

# Topic-BERT

Operations for a response candidate with each conversation utterance

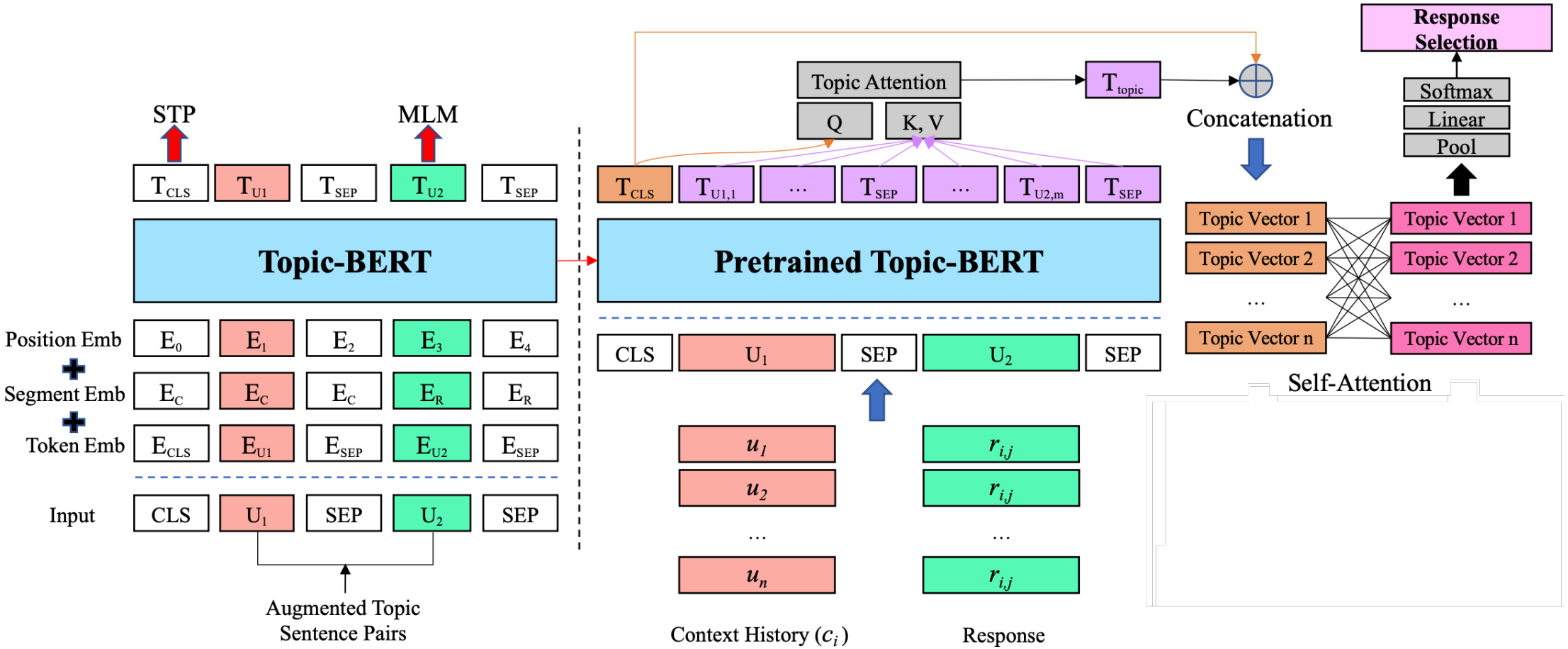





Weishi Wang, et al. *Response Selection for Multi-Party Conversations with Dynamic Topic Tracking*. EMNLP 2020.

# Topic-BERT

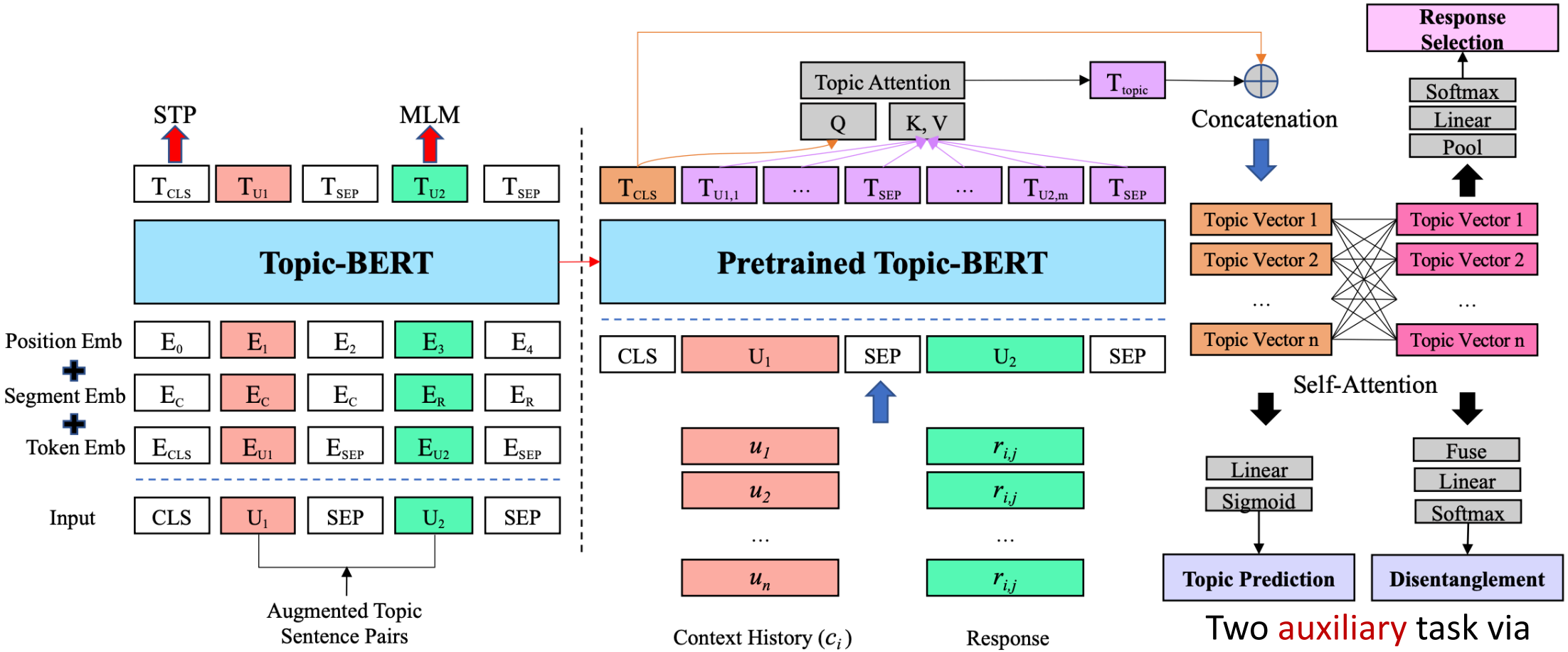
Response selection as the main task






Weishi Wang, et al. *Response Selection for Multi-Party Conversations with Dynamic Topic Tracking*. EMNLP 2020.

# Topic-BERT



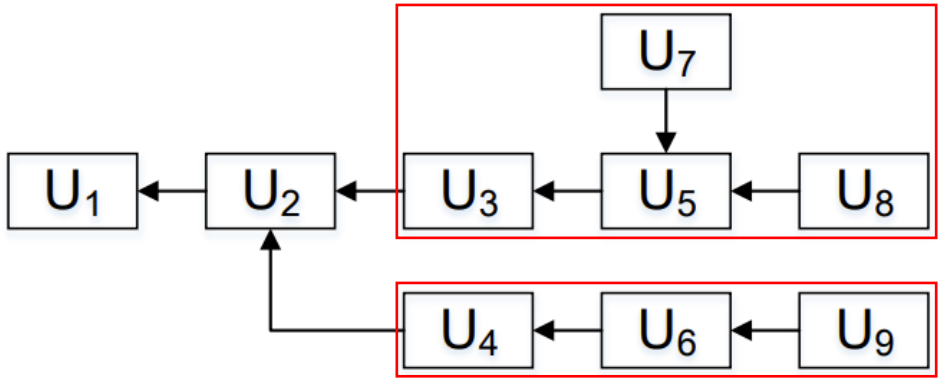

Two auxiliary task via multi-task learning



Weishi Wang, et al. *Response Selection for Multi-Party Conversations with Dynamic Topic Tracking*. EMNLP 2020.

# 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

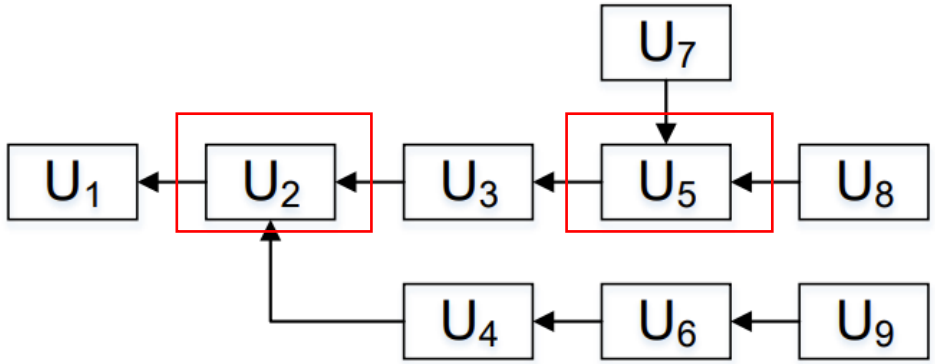



- **Assumption**: the representations of sub-conversations under the same parent node tend to be similar

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




- **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% $R_{10}@1$

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% $R_{10}@1$

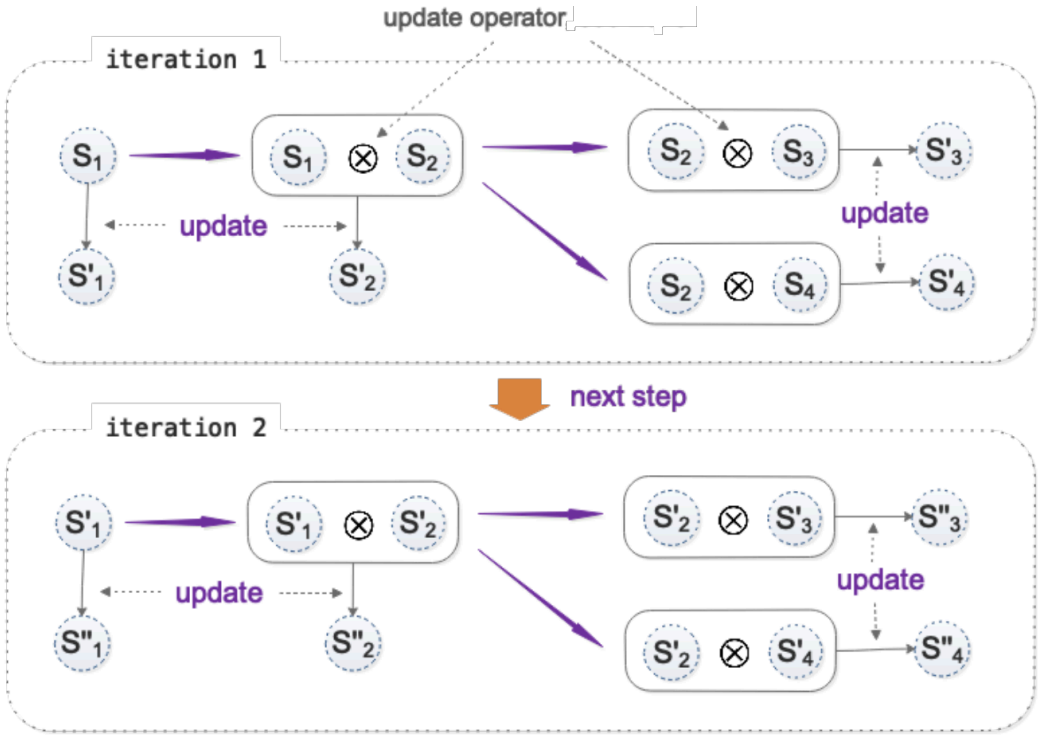
| | Hu et al. (2019) | | Ouchi and Tsuboi (2016) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Len-5 | | Len-10 | | Len-15 | |
| | $R_2@1$ | $R_{10}@1$ | $R_2@1$ | $R_{10}@1$ | $R_2@1$ | $R_{10}@1$ | $R_2@1$ | $R_{10}@1$ |
| DRNN (Ouchi and Tsuboi, 2016) | - | - | 76.07 | 33.62 | 78.16 | 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

# 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

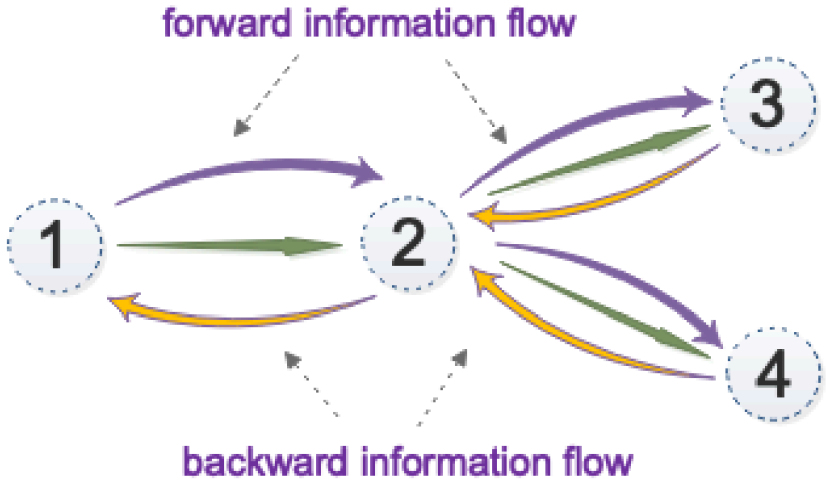
Each utterance (a node in the graph) accepts information from all its connected utterances (nodes) in each iteration

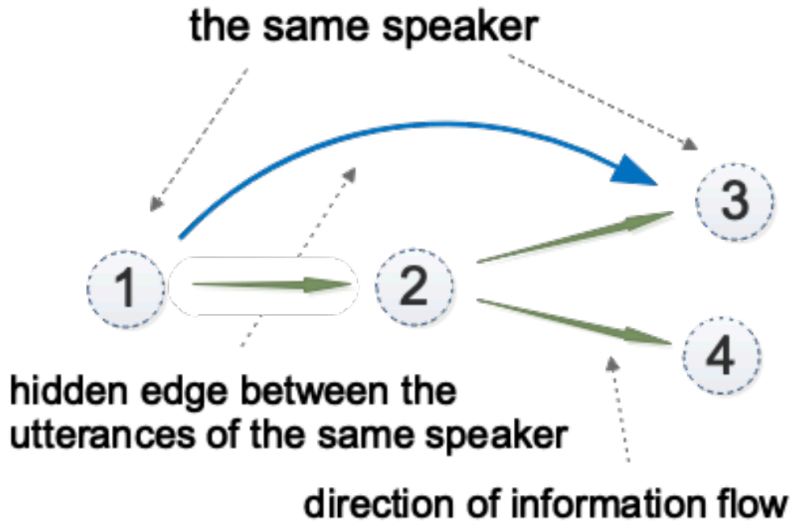
Homogeneous graph composed of only utterances!



Wenpeng Hu, et al. *GSN: A Graph-Structured Network for Multi-Party Dialogues*. IJCAI 2019.

# 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



Wenpeng Hu, et al. *GSN: A Graph-Structured Network for Multi-Party Dialogues*. IJCAI 2019.

# Is a homogeneous graph expressive enough to represent an MPC?

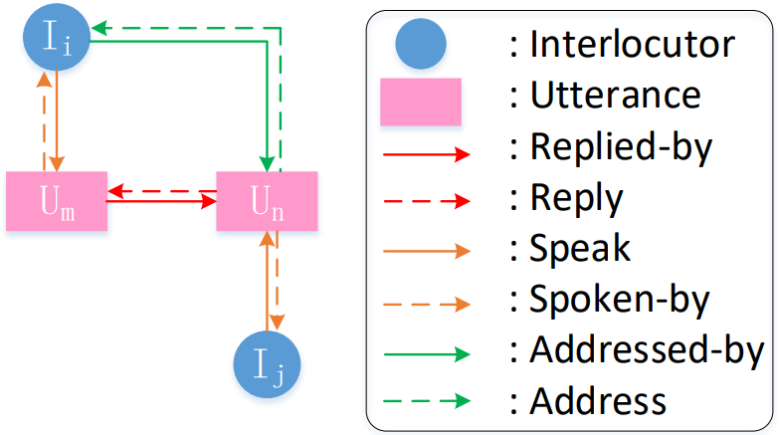
Q1: Are there other sources of information in addition to utterances that should be embraced in a unified graph?

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 → 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

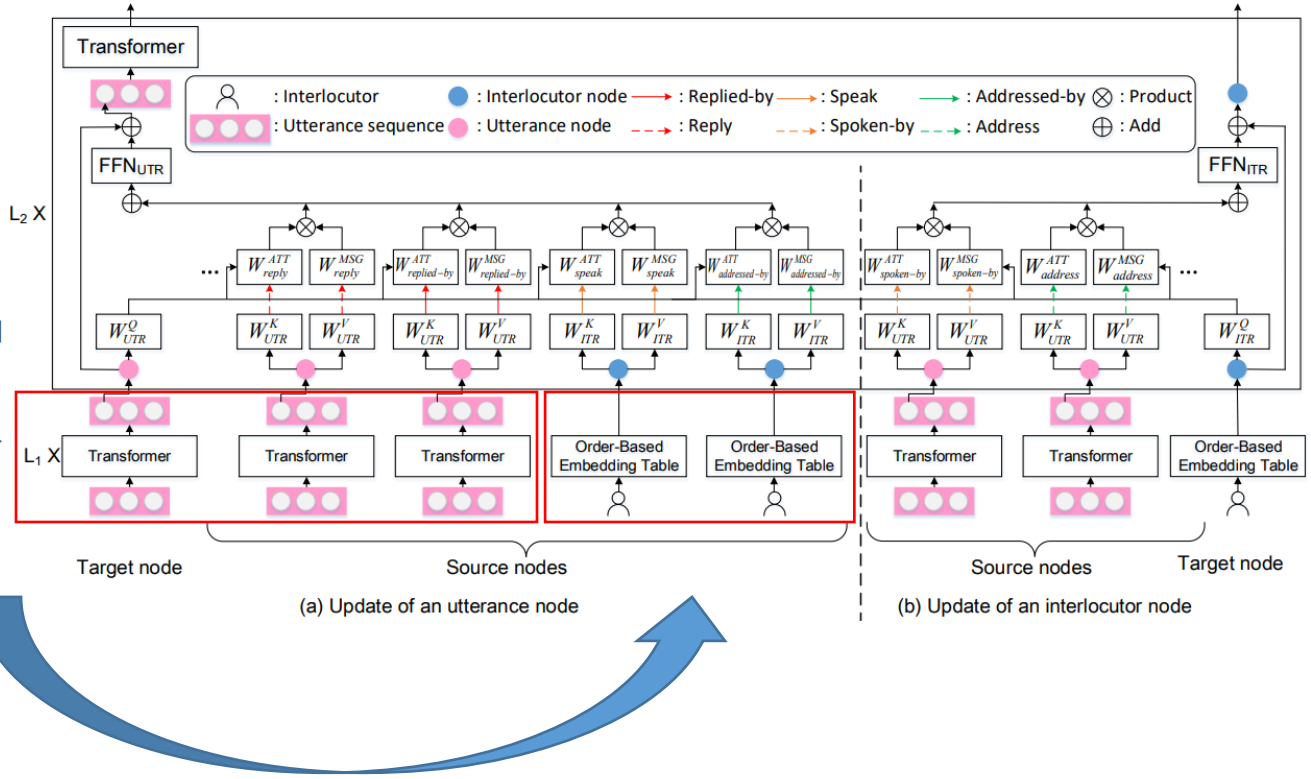





Jia-Chen Gu, et al. *HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations*. ACL 2022.

# HeterMPC: Node Initialization

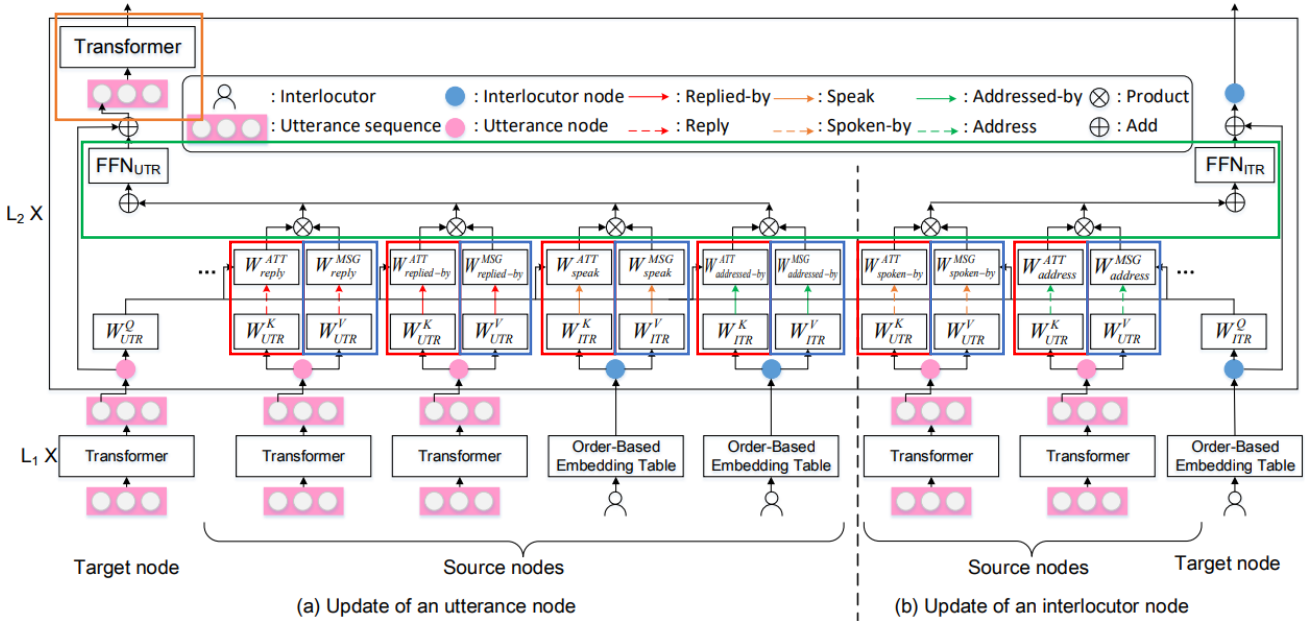
- Each utterance is encoded individually by stacked Transformer encoder layers
- Each interlocutor is directly represented by looking up a position-based interlocutor embedding table



# HeterMPC: Node Updating

Introduce parameters to model heterogeneity via

- 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



Ziniu Hu, et al. *Heterogeneous Graph Transformer*. WWW 2020.

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

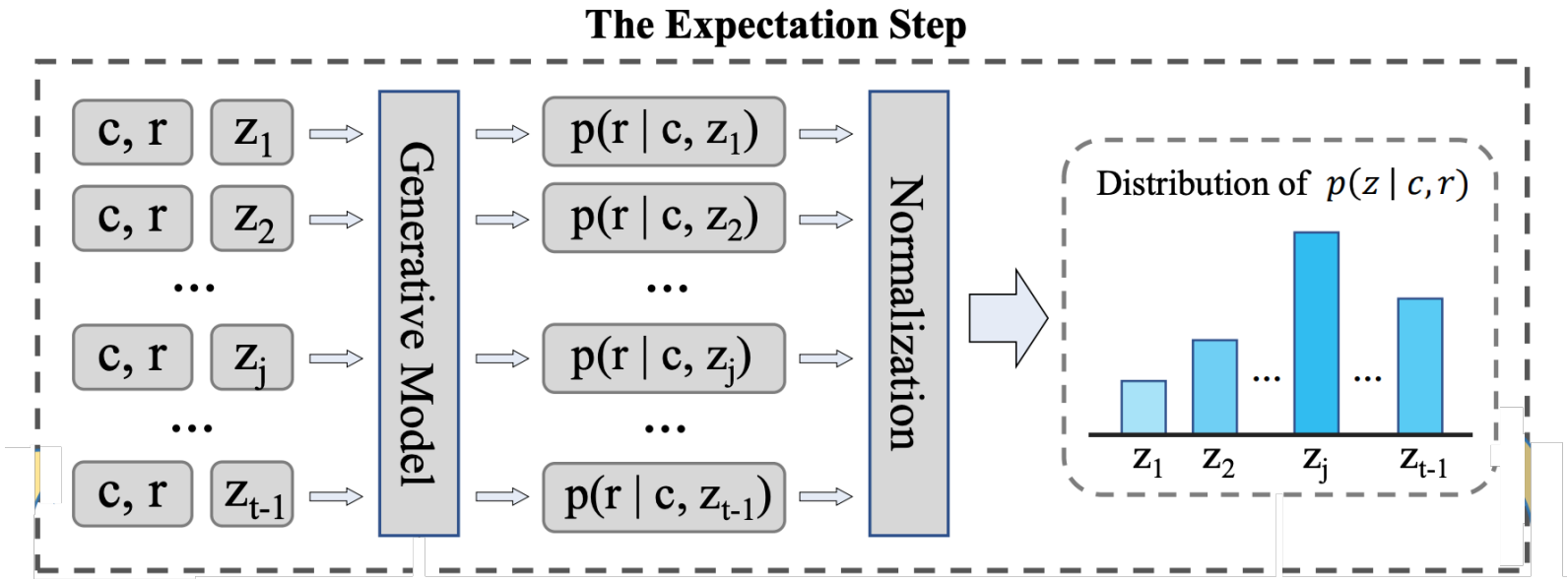
# EM Pre-training

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



Yiyang Li, et al. *EM Pre-training for Multi-party Dialogue Response Generation*. ACL 2023.

# EM Pre-training

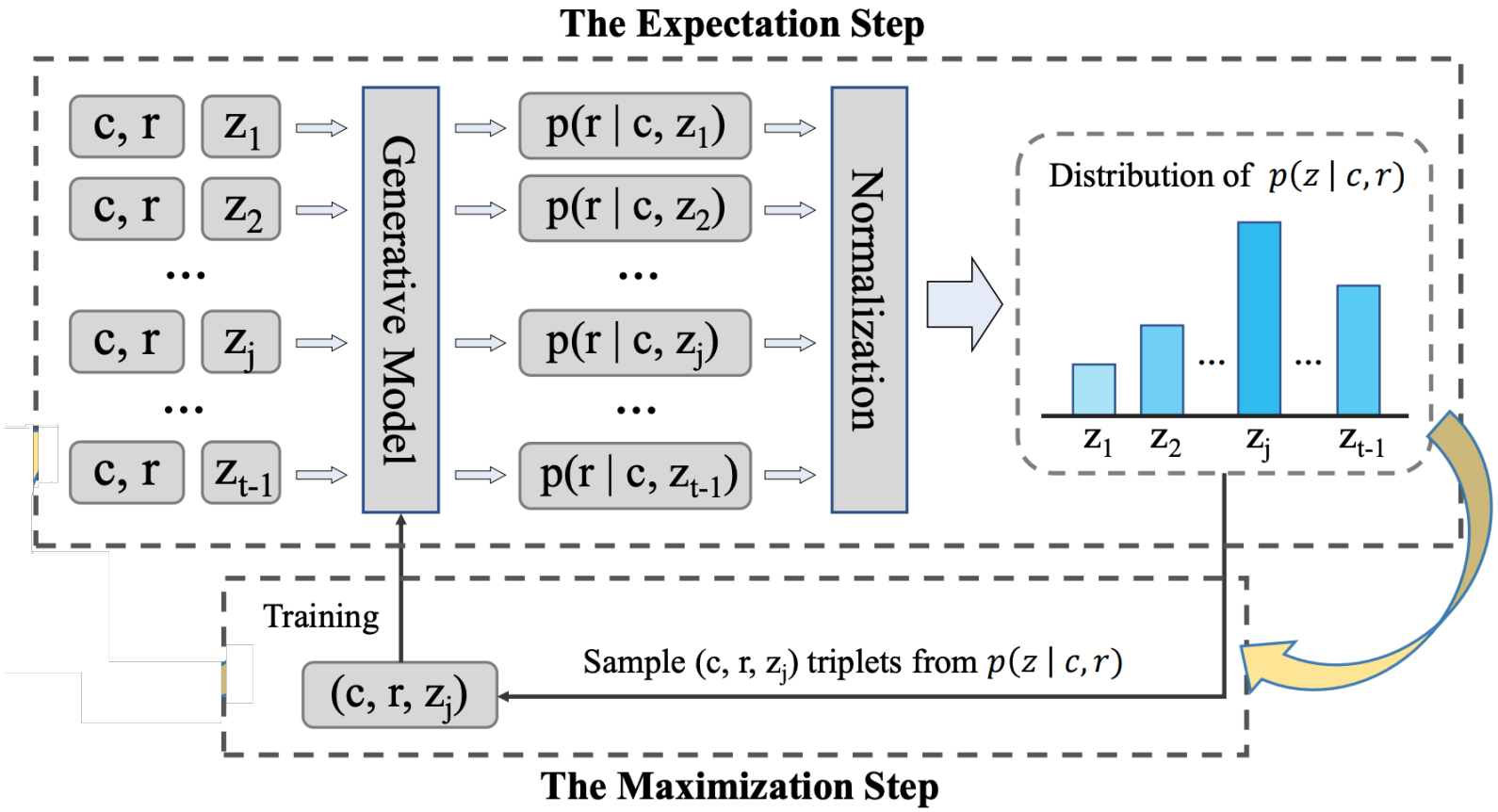



Expectation steps: model the distribution of the latent addressee given the dialogue history and the response



Yiyang Li, et al. *EM Pre-training for Multi-party Dialogue Response Generation*. ACL 2023.

# EM Pre-training

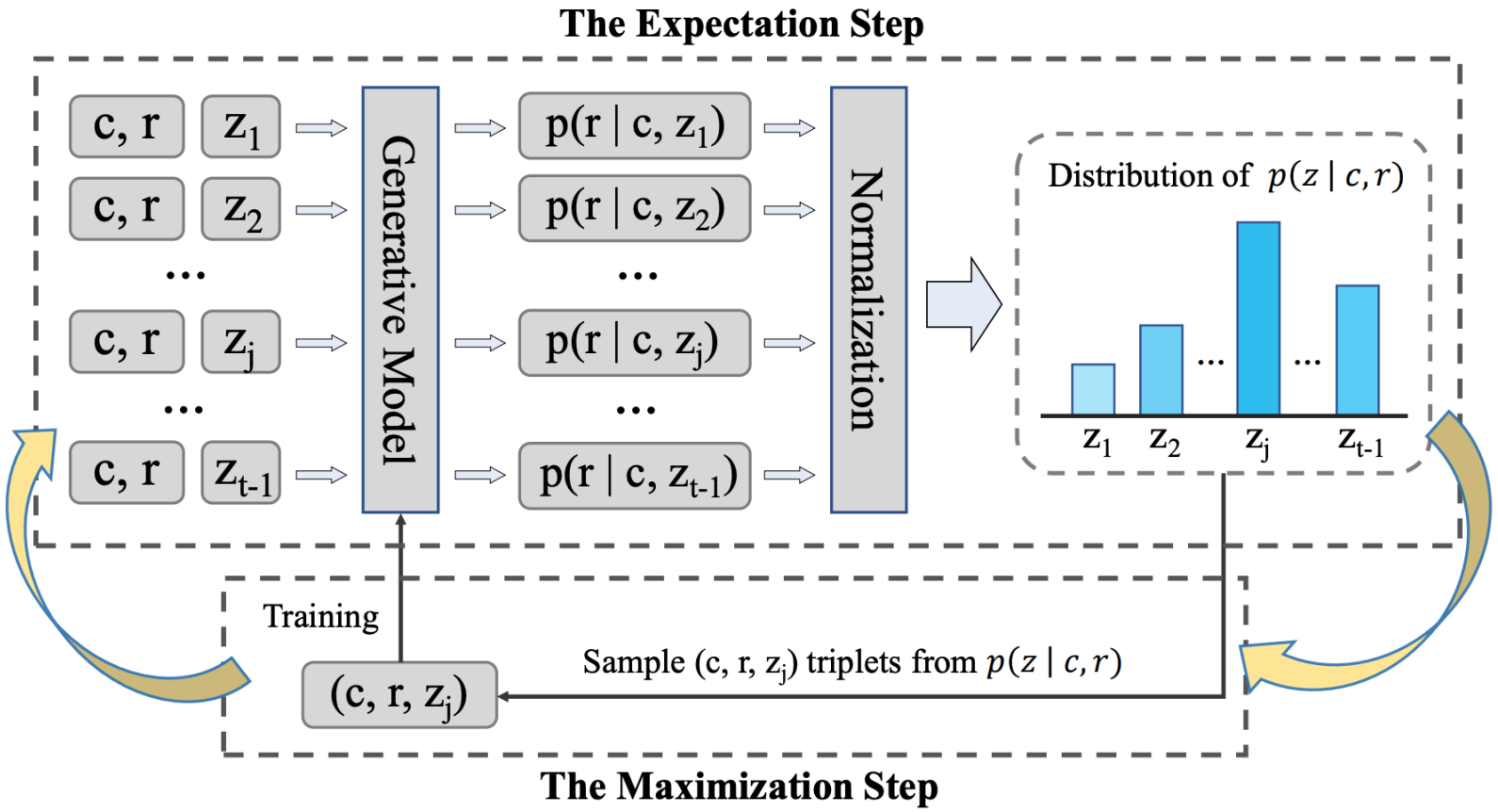



Maximization steps: sample from the addressee distribution and optimize the generative model on these samples



Yiyang Li, et al. *EM Pre-training for Multi-party Dialogue Response Generation*. ACL 2023.

# EM Pre-training




Improve the accuracy of latent addressee distribution and the quality of generative models with more iterations



Yiyang Li, et al. *EM Pre-training for Multi-party Dialogue Response Generation*. ACL 2023.

# 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 5:
# 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

# Challenges


- Reduce the heavy dependency on the necessary addressee labels, while the scarcity of addressee labels is still a common issue in MPC
- Still don’t know yet how to better model the core issues of interlocutor and conversation structure
- Extend to multi-modal MPC, including face and speech interactions, for multi-dimensional agent simulation
- Still lack of high-quality datasets constructed for MPC via data-centric

# Q & A

Thank you for joining us today!

All the materials are at
http://home.ustc.edu.cn/~gujc