Enhanced Speaker-Aware Multi-Party Multi-Turn Dialogue Comprehension

Xinbei Ma †, Zhuosheng Zhang ‡, and Hai Zhao §

Abstract—Multi-party multi-turn dialogue comprehension brings unprecedented challenges in handling complicated scenarios, as the co-occurrence of multiple speakers causes complexity and inconsistency. As a result of the multiple participation, the shift of speaker roles and crisscrossed discourse relations among utterances hinder reading comprehension. Motivated by this, we further integrate the enhancements of speaker-related features for dialogue comprehension performance. This work proposes a novel model with enhancement from both sides of speaker roles and speaker-aware relations. At the token level, we apply a speaker mask for attention, while at the discourse level, we utilize heterogeneous graph networks for comprehensive speaker-aware discourse understanding. Experimental results show that our Enhanced Speaker-Aware method (ESA) helps achieve state-of-the-art performance on the Molweni dataset, as well as significant improvements on the FriendsQA dataset. We find that our method makes steady improvements on stronger backbones. Analysis shows that our model enhances the connections between utterances and their own speakers and captures the speaker-aware discourse relations. Discussions on data features and error cases are presented, and a visualized case is displayed. The findings reveal the importance of speaker-aware signals in dialogue comprehension.

Index Terms—Natural language processing, machine reading comprehension, multi-turn multi-party dialogue, question answering, discourse analysis, graph network.

I. INTRODUCTION

MOTIVATION: Training models to understand dialogue contexts and answer questions has been shown to be even more challenging than common machine reading comprehension (MRC) tasks on plain texts [12], [63]. Dialogues in the multi-party scenario involve multiple utterances announced by three or more speaker roles [38], which is even more intractable than two-party shift [51], [78], [91].

Compared to plain text and two-party dialogues, some inherent features increase the complexity of multi-party scenarios. (i) Each speaker role has an individual intention or speaker manner. This leads to some latent connections between utterances from the same speakers. For example, if a speaker starts a conversation for the purpose of fixing a software bug, then his utterances may all be related to this bug. Or, if a speaker tends to respond to an exposition, then his utterances are more likely to show close attention to prior ones. (ii) Compared to the two-party situation, it is much more difficult to predict the transitions of speaker roles. Each message is available for all participants. Thus, anyone who would like to reply to the message can be the speaker of the next utterance. As a result, continuity is broken due to the presence of shift and crossing dependencies, which are commonplace in a multi-party chat. (iii) There could be a conversation between any two or more speakers within one dialogue history. Hence, interrelations between speaker-aware utterances are much more flexible than plain text or two-party dialogues, whose relations mainly exist in adjacent sentences. Thus, multi-party dialogues appear discourse dependency relations between non-adjacent utterances and lead up to a graphical discourse structure [38], [66]. In other words, it is quite probable that an utterance has a strong connection with some non-adjacent utterance. And that increases the difficulty of dialogue understanding.

Fig. 1 illustrates the challenges of multi-party dialogue MRC, whose corresponding speaker-aware discourse structure of the example dialogue is depicted with different colors indicating different speakers. This example dialogue involves four speakers. The conversation develops as Dr. Willis, NickGarvey and smo help benkong2 with a system error. Along with the context, two relevant questions are annotated and expected to be answered. An extractive span is given as an answer of Question 1, while Question 2 is unanswerable only based on this dialogue.

Prior work: The mainstream work of dialogue MRC commonly adopts Pre-trained Language Models (PrLM) [15], [17] as an encoding module. The pairwise dialogue passage and question are constructed as a whole sequence input to get contextualized, but coarse representation [24], [38], [62]. Recent works have become aware of the characteristics of dialogue passages, taking specific features into account as assistants of dialogue modeling. Speaker information of utterances has been embedded to model two-party shift [24], [46]. Topic information is also used when matching a response with a context [74], [82]. Dialogue Act is utilized to model use intent types [84]. So are emotion [95] and visual information [67]. Whereas, according to our analysis above, speaker-aware features, especially speaker role and speaker-aware relations, deserve strong enhancement. However, there is still room to refine the modeling approaches for speaker-aware features.

Abstract—Multi-party multi-turn dialogue comprehension brings unprecedented challenges in handling complicated scenarios, as the co-occurrence of multiple speakers causes complexity and inconsistency. As a result of the multiple participation, the shift of speaker roles and crisscrossed discourse relations among utterances hinder reading comprehension. Motivated by this, we further integrate the enhancements of speaker-related features for dialogue comprehension performance. This work proposes a novel model with enhancement from both sides of speaker roles and speaker-aware relations. At the token level, we apply a speaker mask for attention, while at the discourse level, we utilize heterogeneous graph networks for comprehensive speaker-aware discourse understanding. Experimental results show that our Enhanced Speaker-Aware method (ESA) helps achieve state-of-the-art performance on the Molweni dataset, as well as significant improvements on the FriendsQA dataset. We find that our method makes steady improvements on stronger backbones. Analysis shows that our model enhances the connections between utterances and their own speakers and captures the speaker-aware discourse relations. Discussions on data features and error cases are presented, and a visualized case is displayed. The findings reveal the importance of speaker-aware signals in dialogue comprehension.

Index Terms—Natural language processing, machine reading comprehension, multi-turn multi-party dialogue, question answering, discourse analysis, graph network.

I. INTRODUCTION

MOTIVATION: Training models to understand dialogue contexts and answer questions has been shown to be even more challenging than common machine reading comprehension (MRC) tasks on plain texts [12], [63]. Dialogues in the multi-party scenario involve multiple utterances announced by three or more speaker roles [38], which is even more intractable than two-party shift [51], [78], [91].

Compared to plain text and two-party dialogues, some inherent features increase the complexity of multi-party scenarios. (i) Each speaker role has an individual intention or speaker manner. This leads to some latent connections between utterances from the same speakers. For example, if a speaker starts a conversation for the purpose of fixing a software bug, then his utterances may all be related to this bug. Or, if a speaker tends to respond to an exposition, then his utterances are more likely to show close attention to prior ones. (ii) Compared to the two-party situation, it is much more difficult to predict the transitions of speaker roles. Each message is available for all participants. Thus, anyone who would like to reply to the message can be the speaker of the next utterance. As a result, continuity is broken due to the presence of shift and crossing dependencies, which are commonplace in a multi-party chat. (iii) There could be a conversation between any two or more speakers within one dialogue history. Hence, interrelations between speaker-aware utterances are much more flexible than plain text or two-party dialogues, whose relations mainly exist in adjacent sentences. Thus, multi-party dialogues appear discourse dependency relations between non-adjacent utterances and lead up to a graphical discourse structure [38], [66]. In other words, it is quite probable that an utterance has a strong connection with some non-adjacent utterance. And that increases the difficulty of dialogue understanding.

Fig. 1 illustrates the challenges of multi-party dialogue MRC, whose corresponding speaker-aware discourse structure of the example dialogue is depicted with different colors indicating different speakers. This example dialogue involves four speakers. The conversation develops as Dr. Willis, NickGarvey and smo help benkong2 with a system error. Along with the context, two relevant questions are annotated and expected to be answered. An extractive span is given as an answer of Question 1, while Question 2 is unanswerable only based on this dialogue.

Prior work: The mainstream work of dialogue MRC commonly adopts Pre-trained Language Models (PrLM) [15], [17] as an encoding module. The pairwise dialogue passage and question are constructed as a whole sequence input to get contextualized, but coarse representation [24], [38], [62]. Recent works have become aware of the characteristics of dialogue passages, taking specific features into account as assistants of dialogue modeling. Speaker information of utterances has been embedded to model two-party shift [24], [46]. Topic information is also used when matching a response with a context [74], [82]. Dialogue Act is utilized to model use intent types [84]. So are emotion [95] and visual information [67]. Whereas, according to our analysis above, speaker-aware features, especially speaker role and speaker-aware relations, deserve strong enhancement. However, there is still room to refine the modeling approaches for speaker-aware features.
Complex speaker role transitions lead to a sophisticated understanding of characters and relationships between utterances and their own speakers and captures the overall speaker role information on the level of tokens, we employ multi-head attention with speaker-aware masks to weigh tokens from different perspectives. Thus, attention is required to be paid to the interrelations among the utterances.

**Contributions:** In this work, we propose an Enhanced Speaker-Aware approach (ESA) to comprehensively model speaker-aware features for MRC. On the basis of PrLM encoding, we employ extended enhancement modules to explicitly model the speaker-aware features. (i) To capture the overall speaker role feature information on the level of tokens, we employ multi-head attention with speaker-aware masks to weigh tokens from different perspectives. Thus, attention is required to be paid to the interrelations among the utterances.

The empirical studies are conducted on two datasets, Molweni [38] and FriendsQA [85]. Both are extractive Question-Answering benchmarks of multi-party dialogues with nontrivial discourse relationships. Experimental results show that the ESA helps our model gain substantial performance improvements over strong baselines and achieve new state-of-the-art performance on Molweni [38] benchmark. Analysis of our model lies in the model architecture aspect and dialogue characters aspect, which shows that our model enhances the connections between utterances and their own speakers and captures the speaker-aware discourse relations. We find that our method works no matter how strong the backbone is, which can be further improved with the extra assistance of dialogue-related pre-training methods. The findings reveal the importance of speaker-aware signals in dialogue comprehension.

The contribution of this work can be summarized in three folds, (i) a pioneer study of decoupling speaker-aware features on dialogue MRC, with significant performance improvement on dialogue QA benchmarks (state-of-the art on Molweni); (ii) an effective approach using discourse-aware graph networks for speaker property and speaker-aware relation modeling; (iii) comprehensive analysis of various perspectives and systematic research on speaker-related features.

II. BACKGROUND AND RELATED WORK

A. Multi-Party Dialogue

1) Speaker-Related Modeling: Inherent characters of dialogue increase the difficulty of comprehension, especially for the more complex multi-party scenario. This has attracted previous dialogue-oriented work to model specific characters and introduce novel tasks. In this article, we focus on speaker-related modeling.

The speaker property plays an important role in dialogue passages. In a dialogue, each message is uttered by a participant and has one or more potential addressees, bridging between speakers. Speaker-aware tasks are facilitated by the publication of the dialogue benchmark [56] where the sender and addressee are annotated for each utterance. Earlier work relies on the RNN-based encoder of dialogue & speaker history and the matching mechanism between the former representation and candidate participants [36], [56], [88]. More recently, MPC-BERT [26] is pre-trained with novel objectives that are designed for interlocutor relationships and utterance semantics modeling. Superior performances are shown on addressee & speaker recognition, and also related response selection [29], [56]. The speaker identity can be extended to the long-term consistent personality descriptions, i.e. persona [9], [89], [94]. Persona provides grounded concepts that can be used to alleviate contradictions and improve factuality [68], [77]. On the other hand, the speaker role is also a research object for personal knowledge analysis, since the personality can be reflected in utterances. Speaker relationship prediction task is improved through modeling speakers and observing their interactions [32], [71]. Speaker interactions lead to reply-to relationships between utterances, which is explicitly modeled as dialogue disentanglement or thread detection [18], [35]. This work regards utterance as the granularity and predicts pair-wise reply-to relationship, which is consistent with interactions between senders and addressees [44], [52].

2) Dialogue MRC: Machine Reading Comprehension is a general concept for the ability of an artificial intelligence system to understand natural language and make responses. The narrow definition of MRC follows the [Passage, Question, Answer] format, i.e., to answer the question according to the passage. The broader definition covers flexible formulations and various tasks like response generation, sentiment analysis, etc. According to recent research trends of dialogue MRC [25], [90], we include
two mainstream tasks, response selection and question answering in this section [12], [16], [62], [89], [85]. These two problems can be unified as \{Passage, Condition, Answer\} format, where \textit{Condition} is the question or response candidates that the answer is conditioned on. For writing convenience, MRC is denoted as \{P, C, A\} in the following.

Dialogue understanding of MRC has been shown much more challenging than the common reading comprehension [2], [38], [66]. Each utterance has an additional property of the speaker role, which breaks the continuity as that in plain non-dialogue texts. The speaker transition causes latent connections between utterances of specific speakers and complex discourse dependencies. Previous work mainly focuses on matching between the dialogue \(P\) and the questions or responses \(C\). Improved variants of attention mechanism are applied to different granularities like token, utterance, or passage level [30], [48]. As PrLMs have shown the impressive capability of language representation, a general way is to employ a PrLM to encode the input sequence, \(\{P, C\}\). In this case, contextualized information is captured through token-level self-attention [38], [43], [61], [96], while features on discourse remain to be modeled. Ignoring higher-level relationships between utterances, such methods stay at sub-optimal performance.

Recently, dialogue-oriented methods have modeled various features for MRC. Speaker-related features, which we concentrate on, have shown significant effects. GSN [29] constructs an utterance-level graph and builds information flow paths for the same speakers, so as to dynamically model the chronological mind changes of speakers. SA-BERT [24] is proposed for two-party dialogue response selection. A speaker embedding is added at token embedding stage. Then, utterances are reorganized based on two speaker roles, \textit{spoken-from speaker} and \textit{spoken-to speaker}. MPC-BERT [26] enhances interlocutor structure at the pre-training stage. Leveraging addresses annotation, self-supervised objectives are designed to improve the interlocutor relations. More explicitly, the speaker identity can be masked and predicted in a self-supervised way [42]. MDFN [46] decouples the contextualized representation into four channels according to the orders of utterances and speakers. Topic relationship prediction is utilized as an auxiliary objective for response selection in a multi-task way [74]. Dialogue topic transition can also assist the segmented utterance modeling [82]. Disentanglement is proved to facilitate the understanding of long chatting records [31], [52]. Integrating knowledge or common sense in dialogue MRC is a more recent trend. Graph network is devised to model token relationships [45], including co-reference knowledge of entities and relation knowledge based on DialogRE [86]. For external knowledge, SKG [61] and PoDS [92] extract a knowledge graph from a knowledge base and fuses graph representations into the attention module. MDKEMN [11] improves domain specialization by pre-training with domain knowledge. Other characters like emotion [95], dialogue act [84], structured summarization [73] prove effectiveness successively. Dialogue adaptive pre-training is another form of dialogue-specific modeling. Continuous pre-training on PrLMs enhances the utterance-level attention [37], [74], [81], or domain features [40].

### B. Discourse Structure Modeling

Discourse parsing focuses on the discourse structure and relationships of texts, whose aim is to predict the relations between discourse units and to discover the discourse structure between those units. Discourse structure has shown benefits to a wide range of NLP tasks, including MRC on multi-party multi-turn dialogue [5], [20], [31], [57], [70], [80].

#### 1) Dialogue Discourse Parsing

Prior existing studies on linguistics-motivated discourse parsing are based on two annotated datasets, Penn Discourse TreeBank (PDTB) [59] or Rhetorical Structure Theory Discourse TreeBank (RST-DT) [54]. PDTB focuses on shallow discourse relations but ignores the overall discourse structure [6], [8], [60], [83]. In contrast, RST is constituency-based, where related adjacent discourse units are merged to form larger units recursively [7], [33], [41], [47], [75], [87]. However, RST only discovers the relations between neighbor discourse units and only allows tree-structure. Specialized for multi-party dialogues, Segmented Discourse Representation Theory (SDRT) [4] alleviates previous problems, as it allows long-distance and graphic relations. Dialogue discourse parsing attracts attention after the first formulation [1] and corpus STAC [5] are proposed. STAC consists of multi-party chats annotated for discourse parsing following SDRT. Later, the multi-party dialogue QA dataset Molweni [38] adopts the same discourse annotation. Now they are two mainstream benchmarks for the dialogue discourse parsing task. Existing research lines can be summarized into three: 1) Graph-based method, which models the relations of each EDU pair and globally decodes for the discourse structure [58], [72]. 2) Sequential method, which predicts relations of one current utterance and incrementally forms the discourse structure [49], [66]. 3) Mixed method, which leverages the distance sensitivity of the sequential method to improve the graph decoding [19].

#### 2) Application on MRC

Discourse structure clarification helps with downstream tasks. GSN [29] adopts an utterance-level graph-structure encoder, while former methods assume the information flows sequentially. In the graph, three types of information flow are modeled, forward & back information flow and the same speaker information. Their utterance information flow is derived from reply-to relations indicated by the symbol, thus, showing discourse relations to some extent but not specific compared to SDRT. HeterMPC [27] modifies the utterance-level graphic-structure encoder into a heterogeneous graph by adding speaker nodes, whose edges model six relations \{reply, replied-by, speak, spoken-by, address, addressed-by\}. DiscoBERT [79] builds an RST discourse graph and a coreference graph on sub-sentence granularity for extractive summarization. Closer to our question-answering task, DADgraph [39] follows Graph Convolutional Network to build graphs for SDRT discourse relations of utterance. Discourse parsing task also can be an auxiliary training objective to improve reading comprehension in a multi-task way [28].

Compared to all the inspiring related studies above, our work is unique and distinguishing. 1) Some existing studies only discuss two-party dialogue or assume dialogue data are two-party [24], [46]. Therefore, the speaker modeling degenerates into a switch. But there is still a gap between two-party dialogues...
and multi-party scenarios. We clarify that our model is designed for the general multi-party dialogue setting, and thus is supposed to be used on all discriminative MRC tasks. But question answering is considered as the end task here because QA task has datasets that are annotated well and QA represents the classic MRC formulation. 2) Different from previous work, utterance relations are regarded as a reflection of speaker transition information. Thus, speaker property and speaker-aware relations are unified as the speaker-aware feature, the primary dialogue characteristic. We decouple the speaker-aware feature by three modules, enhancing speaker role, discourse relationship, and speaker relationship. 3) This work models the speaker-related discourse structure in dialogue MRC to tackle the discourse tangle caused by speaker role transitions. Compared to other graph-based methods [27], [29], [39], our model uses two separate graphs to decouple the two kinds of relations, which avoids information mix and alleviates complexity. Improved performance of empirical studies on two QA benchmarks, Molweni [38] and FriendsQA [85] prove our method effective.

III. METHODOLOGY

We propose an Enhanced Speaker-Aware (ESA) model in this section. As is shown in Fig. 2, ESA follows the paradigm of encoder-decoder, and models speaker-aware features through three extended modules. In general terms, ESA contains a PrLM for encoding, three separated modules to enhance speaker-aware features respectively, and a span extraction layer as a decoder to generate a final answer prediction. The enhancement modules, namely Speaker Masking, Speaker Graph and Discourse Graph, will be introduced in detail, following a formal description of the task.

A. Task Formulation

Supposing we study MRC on a multi-party multi-turn dialogue context \( C \) consisting of \( n \) utterances, which can be represented as \( C = \{U_1, U_2, \ldots, U_n\} \). Each utterance \( U_i \) includes a name identity of the speaker and a sentence by the speaker, denoted by \( U_i = \{S_i, W_i\} \). The sequence \( W_i \) can be denoted as a \( l_i \)-length sequence of words, \( W_i = \{w_1, w_2, \ldots, w_{l_i}\} \). According to this multi-party multi-turn context \( C \), a question \( Q \) is put forward. For this question, the model is expected to find a span from the dialogue context as a correct answer \((p_{\text{start}}, p_{\text{end}})\), or make a decision that this question is impossible to answer only based on the provided dialogue context.

B. Encoding

Following existing work, ESA first encodes the context and question with a PrLM. For better contextualized representations, we concatenate the dialogue context and a question in the form of [CLS] question [SEP] context [SEP]. As we need to process on utterance units, we insert [SEP] token between each pair of adjacent utterances for convenient division. The concatenated sequence is fed into a PrLM, and the output of the PrLM is the initial contextualized representations for each token, denoted as \( H \in \mathbb{R}^{L \times D} \), where \( L \) denotes the input sequence length in tokens, \( D \) denotes the dimension of hidden states.

C. Speaker Masking

As mentioned above, with an additional but inherent speaker feature, words from the same speaker appear to have latent connections. Such words may have an impact on the answer and thus deserve special attention. But self-attention in transformer blocks cannot emphasize this. Hence, our first module is to separately stress speaker property for each token based on speaker transitions of the dialogue passage.

We modify the Multi-Head Self-Attention (MHSA) mechanism with self-designed masks [46], adapting it to multi-party
dialogues. The mask-based MHSA is formulated as follows:

\[
A(Q, K, V, M) = \text{softmax} \left( \frac{Q K^T}{\sqrt{d_k}} + M \right) \cdot V,
\]

\[
\text{head}_i = A \left( H W^Q_i, H W^K_i, H W^V_i, M \right),
\]

\[
\text{MHSA}(H, M) = [\text{head}_1, \text{head}_2, \ldots, \text{head}_N] \cdot W^O,
\]

where \(A\), \(\text{head}_i\), \(Q\), \(K\), \(V\), \(M\) denote the attention, head, query, key, value and mask, \(H\) denotes the original representations from PrLM, and \(W^Q_i, W^K_i, W^V_i, W^O\) are parameters. Operator \([\cdot, \cdot]\) denotes concatenation. Instead of speaking in turn between two people, we have to identify the speaker of each utterance explicitly. In the implementation, we build a vector to embed the speaker labels of each input token, naturally cut by utterance transition. Accordingly, we assign masks for tokens of utterances from the same speaker and different speakers, forming two channels. The two masks are denoted as:

\[
M_1[i, j] = \begin{cases} 0, & S_i = S_j, \\ -\infty, & \text{otherwise}, \end{cases}
\]

\[
M_2[i, j] = \begin{cases} 0, & S_i \neq S_j, \\ -\infty, & \text{otherwise}, \end{cases}
\]

\[
\text{Channel}_1 = \text{MHSA}(H, M_1),
\]

\[
\text{Channel}_2 = \text{MHSA}(H, M_2),
\]

where \(S\) denotes the speaker identity, \(M_1\) and \(M_2\) denote masks of the same speaker and different speakers. These masks can be interpreted as that a negative infinity to eliminate attention we do not need. \(\text{Channel}_1\) contains the decoupled representation of the same speaker while \(\text{Channel}_2\) contains the decoupled representation of the different speakers.

Then we fuse \(\text{Channel}_1\), \(\text{Channel}_2\) and the original contextualized representation \(H\) by a gate-based fusing method [46], which is formulated as:

\[
E_1 = \text{ReLU}(\text{FC}([H, C_1, H - C_1, H \odot C_1])),
\]

\[
E_2 = \text{ReLU}(\text{FC}([H, C_2, H - C_2, H \odot C_2])),
\]

\[
G = \text{Sigmoid}(\text{FC}([E_1, E_2])),
\]

\[
H_C = H \odot C_1 + (1-G) \odot C_2,
\]

where \(C_1\) and \(C_2\) denote the shorthand of the two channels and FC is the shorthand of a fully connected layer. In this case, \(H_C\), the speaker-aware representation is derived and is of the same size as the original contextualized representation \(H\).

D. Graph Modeling

Complicated transitions of speaker roles segment text into separated utterances and break the consistency of passage, thus resulting in intricate interrelations among utterances. From this perspective, utterances are the processing units of discourse-level relationships. Representations are on the granularity of utterance and equally assigned to tokens within the same utterances. We decouple speaker-aware relations into two types: speaker identity and speaker-related discourse dependencies. Note that discourse relationships reflect speaker transitions.

We borrow the graph neural network to construct two separate heterogeneous graphs to avoid information mix or entanglement. To avoid information mix or entanglement, we construct two separate heterogeneous graphs named Speaker Graph and Discourse Graph. There are three kinds of vertices, utterance node, discourse relation node, and context-aware global node. According to the source and target node types, edges are divided into different types and handled separately. The information transformation process follows relational Graph Convolutional Network (r-GCN) [63], where different weights are learned for each type of edge. Speaker graph models the same speaker identity of utterances. Discourse graph models speaker-aware discourse parsing relations, which are resulted from the complex non-adjacent dependencies caused by speaker transitions and thus capture the latent speaker-related information.

1) Speaker Graph: Since the speaker roles in each utterance impact the dialogue development hugely, we build a speaker graph to model relations of utterances based on speaker identity. Specifically, we build an r-GCN and connect utterances from the same speaker, allowing information exchange among statements of identical speakers. This is for the purpose of implicitly modeling connections between the same speakers, e.g., intention or speaking manner. We denote the graph as \(G_s = (V_s, E_s)\), where \(V_s\) is for the set of vertices and \(E_s\) is for the set of edges. First we add vertices \(v_1^s, v_2^s, \ldots, v_n^s\) to represent every single utterance and a special global vertex \(v_{d+1}^s\) for context-level information, denoted as:

\[
V_s = (v_1^s, \ldots, v_n^s, v_{d+1}^s),
\]

where \(n\) is the number of utterances. For each pair of utterances sharing the same speaker, we construct one edge and a reverse edge, which is denoted as \(v_i^s \leftrightarrow v_j^s, S_i = S_j\). Then, we construct a self-directed edge, \(v_i^s \rightarrow v_i^s\), for each vertex and we connect the global vertex to every other vertex, denoted as \(v_{d+1}^s \rightarrow v_i^s\). Fig. 3 illustrates the graph structure of the example dialogue in Fig. 1, with different colors for different kinds of edges.

The contextualized representations of [SEP] tokens are extracted as a simplified pooling. They serve as the original representations of utterance vertices, i.e., the input of the graph network. And the original representations of the global vertex

![Graph Structure](image-url)

**Fig. 3.** Speaker graph of the example dialogue in Fig. 1.
are formed by embedding. The information exchange process can be formulated as:

\[ h_i^{(l+1)} = \sigma \left( \sum_{r \in \mathbb{R}} \sum_{j \in N_r^i} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)} \right), \]

(5)

where \( \mathbb{R} \) denotes the set of relations with other vertices. \( N_r^i \) denotes the set of neighbours of vertex \( v_i \), which are connected to \( v_i \) through relation \( r \), and \( c_{i,r} \) is the element number of \( N_r^i \) used for normalization. \( W_r^{(l)} \) and \( W_0^{(l)} \) are parameter matrices of layer \( l \). \( \sigma \) is activated function, which in our implementation is ReLU [3], [23], [55]. After information exchange with neighbour nodes, we get the vectors of each utterance, containing speaker interrelation information. After \( L_s \) layers, we get \( H_{S}^{L_{s}} \in \mathbb{R}^{(n+1) \times D} \) as the last-layer output of the graph. As this modelling is on the granularity of utterance while the prediction is on tokens, a size extension is needed. Based on the intuition that each token inside the same utterance shares the same speaker information, we equally expand \( H_{S}^{L_{s}} \) to the same dimension of \( H \) for later fusion, which is denoted as \( H_{S} \in \mathbb{R}^{L \times D} \). The extension is illustrated in Fig. 4.

2) Discourse Graph: Discourse relations exist in speaker-aware utterances and implicitly contain speaker-related information. All the considered discourse relationships are listed in Table I, along with some necessary clarifications of meanings [38], [85]. In parallel to the speaker graph, we build a graph according to the annotated discourse relations to connect relevant utterance pairs. The pre-processing includes two steps. First, we assign a label for every considered relationship. Second, we reserve each relationship in the form of \( \) (first utterance, second utterance, relation label).

Then the graph is constructed according to the simplified representations of relations. We denote the graph as \( G_{d} = (V_{d}, E_{d}) \), where \( V_{d} \) is for the set of vertices and \( E_{d} \) is for the set of edges. The following kinds of vertices are constructed into the graph: utterance vertices for each utterance, relation vertices for each existing relation, and a global vertex to represent the dialogue-level information. Vertices can be denoted as:

\[ V_{d} = (v_{d}^{1}, \ldots, v_{d}^{n}, v_{d}^{n+1}, v_{d}^{n+2}, \ldots, v_{d}^{n+n_{r}+1}), \]

(6)

where \( n \) is the number of utterances and \( n_r \) is the number of corresponding relations. In terms of \( E_{d} \), for each relation \( (v_{d}^{i}, v_{d}^{j}, r_m) \), we construct oriented edges \( v_{d}^{i} \rightarrow r_m \) and \( r_m \rightarrow v_{d}^{j} \), and also reverse oriented edges \( r_m \rightarrow v_{d}^{i} \) and \( v_{d}^{j} \rightarrow r_m \). As the same as Speaker Graph, we add a self-directed edge \( v_{d}^{i} \rightarrow v_{d}^{j} \) to each vertice. And for each vertice representing utterances or relations, a global vertice-directed edge \( v_{d}^{n+1} \rightarrow v_{d}^{i}, i \neq n + 1 \) is added. An example is shown in Fig. 5.

Similar to the Speaker Graph, the input vector of utterance vertices are the contextualized representations of [SEP] token. The original representations of relation vertices and the global vertice are formed by an embedding layer. The formulation of message-passing is the same as the Speaker Graph, except that the relationship set \( \mathbb{R} \) contains more kinds of relations, as shown in Table I. After the message-passing between related vertices, we get the vectors of each utterance containing speaker-aware discourse structure as output. The output of the last layer of discourse graph is denoted as \( H_{G}^{L_{s}} \in \mathbb{R}^{(n+n_{r}+1) \times D} \), we keep the vectors for utterances \( H_{G}^{L_{s}}[0 : n] \) and perform the same extension shown in Fig. 4, then we get \( H_{G} \in \mathbb{R}^{L \times D} \).

E. Fusing and Predicting

Decoupled channels \( H_{C}, H_{S}, H_{G} \) from the aforementioned three modules are fused for final answer prediction. We concatenate \( H_{C}, H_{S}, H_{G} \) and \( H \) together to obtain the overall speaker-enhanced contextualized representations:

\[ P = [H_{C}, H_{S}, H_{G}, H]. \]

(7)

Following the standard process for span-based MRC [17], [22], [93], the representation \( P \) is fed to a fully connected layer, calculating the probability distribution of the start position and end position for an answer span. Cross-entropy function is used as the training object to minimize. And our loss function is the average of the cross-entropy of the start position, end position, and a binary cross-entropy of the possibility to be answerable.

IV. EXPERIMENTS

A. Datasets

ESA is evaluated on two datasets of multi-party multi-turn dialogue MRC, namely Molweni [38] and FriendsQA [85].

Fig. 4. Extension of output of speaker graph.

Fig. 5. Discourse graph of the example dialogue in Fig. 1.

### Table I

| Relationship names (meanings) |
|-------------------------------|
| Comment, Clarification Question, QAP (Question Answering Pair), Continuation, Acknowledgement, Result, Elaboration, Correction, Q.Elab (Utter) is a question and Utter2 elaborates Utter1), Contrast, Conditional, Background, Narration, Alteration, Parallel |

**Discourse Relations**
TABLE II
EXPERIMENTAL RESULTS ON THE TEST SET OF MOLWENI AND FRIENDSQA.
ALL RESULTS ARE FROM OUR IMPLEMENTATIONS EXCEPT PUBLIC BASELINES
FOR REFERENCE. "OUR BASELINES" ARE OUR REPRODUCTIONS OF THE PUBLIC BASELINES

| Model                  | Molweni EM | Molweni F1 | FriendsQA EM | FriendsQA F1 |
|------------------------|------------|------------|--------------|--------------|
| **BERT_{base}**        |            |            |              |              |
| Public Baseline [38]   | 45.3       | 58.0       | 45.2         | -            |
| Our Baseline           | 45.7       | 58.8       | 45.1         | 60.1         |
| +Speaker Embedding [24] | 47.1       | 60.5       | 45.0         | 60.1         |
| +MDFN [46]             | 47.1       | 60.5       | 45.0         | 60.1         |
| +ESA                   | 47.1       | 60.5       | 45.0         | 60.1         |

| **BERT_{large}**       |            |            |              |              |
| Public Baseline [38]   | 51.8       | 65.5       | -            | -            |
| Our Baseline           | 52.0       | 65.6       | 47.3         | 63.3         |
| +Speaker Embedding [24] | 52.4       | 65.7       | 46.8         | 63.3         |
| +MDFN [46]             | 52.4       | 65.7       | 46.8         | 63.3         |
| +ESA                   | 52.4       | 65.7       | 46.8         | 63.3         |

| **BERT_{wwm}**         |            |            |              |              |
| Public Baseline [38]   | 54.7       | 67.6       | -            | -            |
| Our Baseline           | 53.9       | 67.5       | 50.1         | 66.2         |
| +Speaker Embedding [24] | 56.0       | 68.3       | 49.2         | 65.9         |
| +MDFN [46]             | 55.8       | 68.7       | 50.4         | 66.2         |
| +ESA                   | 56.0       | 69.1       | 52.1         | 68.0         |

| **ELECTRA_{large}**    |            |            |              |              |
| Public Baseline [38]   | 57.3       | 70.4       | 56.8         | 74.0         |
| Our Baseline           | 57.9       | 72.1       | 56.7         | 74.0         |
| +Speaker Embedding [24] | 57.9       | 72.1       | 57.8         | 75.2         |
| +MDFN [46]             | 57.9       | 72.1       | 57.8         | 75.2         |
| +ESA                   | 58.6       | 72.2       | 58.7         | 75.4         |

The bold values indicate the best result.

1) Molweni: Molweni dataset [38] comprises 10,000 multi-party multi-turn dialogues. On average, each dialogue context contains 8.82 utterances from 3.51 speaker roles. The dialogues derive from chatting records in Ubuntu Chat Corpus [50]. On the raw dataset, the following annotations are explicitly made, making Molweni an ideal evaluation benchmark for our research: (i) Answerable and unanswerable extractive questions and golden or reasonable answers. (ii) Elementary Discourse Units (EDUs) cut on the granularity of utterance, including the utterance text and the speaker. (iii) Discourse relationships for each dialogue passage, reflecting interrelations between utterances as Table I.

2) FriendsQA: To verify the generality, we also evaluate our model on FriendsQA [85]. FriendsQA is a challenging multi-party multi-turn dialogue dataset including 1,222 human-to-human conversations from the TV show Friends. 10,610 answerable extractive questions are annotated. Dialogues are segmented into speaker-aware utterances as well. Different from Molweni, two factors may even increase the difficulty of MRC. (i) Based on the lines of TV shows, FriendsQA is open-domain rather than in-domain. (ii) No utterance relationships are provided. Thus we borrow the tool from [66] to annotate discourse relations, which may have a potential bias from the ground truth.

B. Baseline

Following existing work [24], [38], [46], we use BERT as a backbone, where the contextualized $H$ output is used for span extraction directly. We also apply BERT_{large}, BERT_{whole word masking} (BERT_{wwm}) and ELECTRA_{large}\(^1\) as baselines to see if the advance of our method still holds on top of the stronger PrLMs.

We compare our ESA with closely related speaker-aware methods [24], [46] in Table II. Please note that their existing evaluations [24], [46] are conducted on response selection tasks on two-party datasets or datasets without explicit speaker annotations [51], [91]. So we adjust their speaker encoding and end layers for implementations on the QA task of the multi-party scenario. In addition, there are some loosely related studies that also deploy graph networks or speaker features in other ways or for other tasks, but can be applied to our datasets. We present these methods in Table III and show the superior performance of ESA. Baselines shown in table II and III include:

1) Public Baseline. Official results from original papers of datasets [38], [85].
2) Speaker Embedding [24] and MDFN [46]. The methods are described above. Their speaker encoding can be adapted to multi-party dialogue by adding a speaker indicator. The utterance channels of MDFN are kept. And the last layers are changed for span prediction.
3) BiDaF [64]. It models the bidirectional attention flows, i.e., Query to Context & Context to Query, to obtain interactive query-aware context representation.
4) DocQA [13]. It is an attention-based paragraph-level QA method and can scale to the multi-document scenario. The best-matched paragraph is selected as input to DocQA.
5) DialogueRNN [52]. DialogueRNN deploys a GRU to model dialogue states of each speaker respectively for context comprehension enhancement.
6) DialogueGCN [21]. A directed r-GCN network is constructed for dialogue understanding. The vertices represent utterances, and the edges are built between

\(^1\)The weights are \{bert-base/large-uncased, bert-large-uncased-whole-word-masking, electra-large-discriminator\}, available in the Transformers. https://github.com/huggingface/transformers

TABLE III
RELATED STUDIES OF MULTI-PARTY MRC ON THE SAME BENCHMARKS

| Model                  | EM     | F1     |
|------------------------|--------|--------|
| **BERT_{base}** Molweni |        |        |
| Public Baseline [38]   | 45.3   | 58.0   |
| BiDaF [39, 65]         | 22.9   | 39.8   |
| DocQA [14, 39]         | 42.5   | 56.0   |
| DialogueRNN [39, 53]   | 45.4   | 60.9   |
| DialogueGCN [21, 39]   | 45.7   | 61.0   |
| DADgraph [39]          | 46.5   | 61.5   |
| Model of Li and Zhao [42] | 49.2   | 64.0   |
| ESA (Our Model)        | 49.7   | 64.4   |

| **BERT_{large} FriendsQA** |        |        |
| Public Baseline          | 45.2   | -      |
| ULM+UOP [37]             | 46.8   | 63.1   |
| DADgraph [39]            | 46.5   | 61.5   |
| Model of Liu et. al. [45] | 46.4   | 64.3   |
| Model of Li and Zhao [42] | 46.9   | 63.9   |
| ESA (Our Model)          | 47.0   | 63.0   |

The bold values indicate the best result.
utterances within a fixed-length context window. Edges are classified according to different speakers and temporal directions. Thus, the information transformation of r-GCN models speaker dependencies and temporal dependencies.

7) DADgraph [39]. DADgraph constructs an r-GCN for discourse structure, where the vertices represent utterances and edges represent SDRT discourse relationships. Different types of discourse dependencies are convoluted separately in information transformation.

8) ULM+UOP [37]. This work proposes two self-supervised objectives for hierarchical dialogue understanding. Utterance-level Masked LM (ULM) models utterance-level granularity. Utterance Order Prediction (UOP) is for learning attention among utterances.

9) Model of Liu et al. [45]. An r-GCN network is deployed to integrate internal knowledge. The vertices are words in a dialogue, while the edges represent co-reference knowledge and semantic relation knowledge.

10) Model of Li and Zhao [42]. This work proposes a multi-task method with two auxiliary self-supervised training objectives, locating the utterance that contains the answer and predicting the speaker identity.

### C. Setup

Our implementations are based on Transformers [76]. Exact match (EM) and F1 Score are the metrics to measure performance. Our fine-tuning employs AdamW [50] as the optimizer. The learning rate is set to \{3e-5, 5e-5, 4e-6\}. The epoch number is set to \{2, 4\}. The batch size is set to \{8, 16, 32\}. The maximum input length is set to 348. More specifically, the settings of our best model (58.6/72.6 in Table VI) are as follows: 3e-5 for learning rate, 4 for epoch numbers, 32 for batch size. Besides, we set weight decay as 0.01, warm-up step as 100. The code and more implementation details are available\(^2\).

### D. Results

Table II shows the results of our experiments. The empirical results show that our ESA outperforms all baselines on both Molweni and FriendsQA. We can see that our model helps effectively capture speaker role information and speaker-aware discourse structure information and then strengthens the ability of multi-party multi-turn MRC. Table III shows the comparison of ESA with existing works. On Molweni, our ESA achieves state-of-the-art. On FriendsQA, ESA makes the best EM, while improved but modest F1. This may indicate further exploration of the open-domain data and discourse parsing method.

### V. Analysis

#### A. Ablation Study

Since our speaker-aware enhancement includes three separate modules, we perform an ablation study to verify the contributions of the three modules individually. Respectively, we add each aforementioned module and train them under the same hyper-parameters. As shown in Table IV, experimental results indicate that each module plays an effective part in the whole model, and the Speaker Masking module contributes the most. The reason we infer is that the prediction of the answer is token-level distribution, i.e., predict the possible token as start or end position. This may lead to a little advantage of the token-level masking, while the discourse-level graphs still show improvement.

#### B. Graph Structure

1) Construction Methods: We further implement variations of graph construction approaches as an optimization of the graph modules. (i) Considering that the types of discourse relationships can be sparse within one short dialogue, it may make sense to ignore their types. From this perspective, utterances can be connected by the same kind of edges as long as they have some discourse relations. We call such a graph as *Average Graph*. (ii) In our ESA, for the decoupling purpose, we build two separate graphs for speaker relationships and discourse relationships. Separate graphs can make the representations to be more self-informative and less mixed. As a comparison, we combine the two graphs as one. All edges are added within one graph, which we call *Combined Graph*. (iii) As graphs regard utterances as language units, the utterance-level enhancement also deserves attention. As shown in Fig. 4, an utterance representation is extended to all tokens within the same utterance. For comparison, we directly use the initialization of graph nodes, i.e., the \([SEP]\) vectors, to perform this extension. This model is called *UtterEnhancement*.

Table V presents the results. We can see that ignoring the types of discourse relationship type causes performance loss

---

\(^2\)https://github.com/xbmxb/ESA

---

### Table IV

| Model            | EM   | F1   |
|------------------|------|------|
| BERT<sub>base</sub> | 45.3 | 58.0 |
| +Speaker Masking | 49.6 | 63.4 |
| +Discourse Graph | 49.0 | 63.3 |
| +ESA             | 49.7 | 64.4 |
| BERT<sub>large</sub> | 51.8 | 65.5 |
| +Speaker Masking | 52.7 | 65.8 |
| +Discourse Graph | 52.7 | 66.0 |
| +ESA             | 52.9 | 66.9 |
| BERT<sub>base</sub> | 53.9 | 67.5 |
| + Speaker Masking | 55.8 | 68.7 |
| +Discourse Graph | 55.2 | 68.3 |
| +ESA             | 56.0 | 69.1 |
| ELECTRA<sub>large</sub> | 57.3 | 70.4 |
| +Speaker Masking | 57.9 | 71.0 |
| +Discourse Graph | 57.6 | 72.1 |
| +ESA             | 58.4 | 71.8 |
| +ESA             | 58.6 | 72.2 |

The bold values indicate the best result.
TABLE V

| Model                  | EM  | F1  |
|------------------------|-----|-----|
| **BERTbase**           | 45.3| 58.0|
| **Our Baseline**       | 45.7| 58.8|
| w/ Average Graph       | 48.9| (5.3)| 63.8| (5.0)|
| w/ Combined Graph      | 48.6| (4.9)| 63.3| (4.5)|
| w/ UtterEnhancement    | 50.0| (4.3)| 63.4| (4.6)|
| ESA                    | 49.7| (4.0)| 64.4| (5.6)|

**ELECTRAlarge**

| Model                  | EM  | F1  |
|------------------------|-----|-----|
| **Our Baseline**       | 57.3| 70.4|
| w/ Average Graph       | 57.4| (1.0)| 71.7| (1.3)|
| w/ Combined Graph      | 58.4| (1.1)| 71.8| (1.4)|
| w/ UtterEnhancement    | 57.4| (1.0)| 71.3| (0.9)|
| ESA                    | 58.6| (1.3)| 72.2| (1.8)|

The bold values indicate the best result.

TABLE VI

RESULT OF DIFFERENT GRAPH LAYERS

| Model                  | EM  | F1  |
|------------------------|-----|-----|
| **BERTbase**           | 45.3| 58.0|
| **Our Baseline**       | 45.7| 58.8|
| w/ 1layer DG + 1layer SG | 49.7| (4.0)| 64.4| (5.6)|
| w/ 2layer DG + 1layer SG | 49.9| (4.2)| 64.2| (5.4)|
| w/ 3layer DG + 3layer SG | 50.1| (4.4)| 64.3| (5.5)|
| w/ 3layer DG + 1layer SG | 50.1| (4.4)| 64.5| (5.7)|

**ELECTRAlarge**

| Model                  | EM  | F1  |
|------------------------|-----|-----|
| **Our Baseline**       | 57.3| 70.4|
| w/ 1layer DG + 1layer SG | 58.6| (1.3)| 72.2| (1.8)|
| w/ 2layer DG + 1layer SG | 58.6| (1.3)| 72.6| (2.2)|
| w/ 3layer DG + 3layer SG | 58.2| (0.9)| 71.5| (1.1)|
| w/ 3layer DG + 1layer SG | 58.0| (0.7)| 71.5| (1.1)|

The bold values indicate the best result.

of Average Graph. The decreased scores of Combined Graph verify the decoupling of two separate graphs. Whereas, UtterEnhancement gives comparable EM but lower F1. This shows the effectiveness of utterance-level enhancement while may suggest a more powerful representation of each utterance.

2) Graph Layers: As an extension of the analysis of graph structure, we also try different numbers of layers of our proposed graphs. We change the number of layers of the discourse graph and speaker graph simultaneously or respectively. Layer number settings with comparable scores are presented in table VI. The results show deeper r-GCN gives small improvements, which makes sense because the first hop of edges is the relationships that we actively model.

C. Speaker Ablation

To discuss the significance of speaker roles in dialogue modeling, we conduct a straightforward ablation experiment for speaker name annotations. We ablate the texts of speaker names in two ways. (i) We simply delete speaker names in input tokens. In formulation, $S_i$ of each $U_i$ is deleted; thus the input becomes concatenated $W_i$. This is the same situation as datasets without speaker annotations. (ii) Different speaker names are replaced by an identical token, used as a placeholder. This modification can be formulated as $U_i = \{S, W_i\}$. Here we use the word speaker as the placeholder. Speaker ablation is applied to both baseline and our model. Please note that this ablation only manipulates speaker names in contexts. Speaker features are reserved for the three enhancement modules in ESA. [SEP] tokens between utterances are kept as well.

From Table VII we can see that (i) deletion of speakers in context harms both baseline and our model by a quite large margin. This indicates the speaker-name text is significant for dialogue context understanding; (ii) a unified placeholder also shows heavy disadvantages for both baseline and our model, but slightly more moderate than (i).

D. Dialogue-Related Pre-Training

As the unique nature of dialogue passages has drawn attention in the stage of pre-training, studies have made efforts to seek a balance between the generality of natural language and the characteristics of dialogues. Here we analyze the performance of the dialogue-related pre-trained model. To avoid potential data leakage, models that are pre-trained on Ubuntu datasets (e.g., Ubuntu [51], DSTC [34]) are not taken into account. With such purpose, we choose Dialog-PrLM [81] to smooth the transfer between PrLM and dialogue MRC. By manual construction on Wikipedia data, this model is pre-trained on three self-supervised tasks, Insertion, Deletion, and Replacement, simulating the shift or consistency on utterance level. The pre-trained model is applied to the baseline and our model, and Table VIII shows the results. The pre-training shows non-trivial improvement on the baseline while less on our model. This indicates that, on the one hand, these pre-training objectives [81] are verified to be solid; on the other hand, extra enhancements of
our model may have overlapped with such further pre-training on dialogue characters.

### E. Effect of Data Features

Data features can affect the performance more or less. For questions, we consider the interrogative word, while for dialogue contexts, we study the speaker or utterance numbers. The data for analysis is the Molweni test set as is fully annotated. The predictions are from models with ELECTRA\textsubscript{large} as the backbone.

1) **Interrogative Word:** The interrogative words are directional indications for a question because they suggest attention to different components. For example, the answer to *Where*-type question may focus on a phrase following a preposition like *in*. Our study clusters five main types of interrogative words, *Who*, *How*, *Why*, *Where*, *What*, as there are much fewer examples of others, like *Do/Does*, *Be*, *Which*, *etc.* *Whose*-type questions are included in the type *Who* for the similar purpose. Table IX shows the scores of each question type from ELECTRA\textsubscript{large} baseline and our ESA, from which it can be inferred that: (i) For both models, type *How* and *Why* questions are more difficult, and *How*-type questions are easier. This is because *How/Why* questions require longer descriptions while the answers to *Who* questions are entities and often limited within speaker names. (ii) ESA improves QA performance of all types of questions. *Who* question scores improve the most (10.2 EM and 10.1 F1). This type of QA needs a model to locate the relative utterance, then gives the speaker, which directly shows the effects of our utterance-level graphs and the masked attention of words within an utterance. The benefit of utterance relations mainly shows in *How* type.

2) **Speaker and Utterance Numbers:** Fig. 6 shows scores of questions that contain different speaker and utterance numbers. (i) For speaker numbers, ESA makes progress except for 5/6-speaker dialogues. This is because, in the test set, the average utterance numbers of 5/6-speaker dialogues are 7 and 7.2, which means that most speakers only have one utterance. Thus, the speaker-aware modeling methods degenerate, e.g., the edges of the speaker graph may become self-loops. It suggests that excessive sparsity hurts performance. (ii) For utterance numbers, ESA robustly improves the performance. On longer dialogues with 13 utterances, ESA still gains 3.4 on EM and 2.2 on F1. This analysis shows that ESA makes steady improvements on various dialogue contexts, but is influenced by the combination of speaker and utterance numbers. Sparse utterances of different speakers hurt speaker-aware modeling.

### F. Prediction Analysis

To intuitively show how our model improves the ability of MRC on multi-party multi-turn dialogues, we analyze the predictions of the test set from both the baseline and our ESA method. Questions that get wrong answers from the baseline but get good answers from ESA are selected. There are 146 samples whose ESA F1 is higher than baseline F1 by a larger gap of 0.9.

Like Section V-E1, we first take a closer look at the question type. Fig. 7 shows the proportion change between the test set and selected ones. The proportion of questions introduced by
Fig. 8. Selected cases where baseline model fails (Prediction1) but our model gives gold answers (Prediction2). Related segments of dialogues are presented for illustration.

Example 1
Question: Who puts au on all except for security?
Gold answer: lightbright
Prediction1: lightbright
Prediction2: timfrost

Example 2
Question: why permission is denied?
Gold answer: because it’s owned by root
Prediction1: because it’s owned by root
Prediction2: cant save the file

Example 3
Question: How do qkslwrvolf get the FILEPATH manager to use a different icon?
Gold answer: edit the icon theme
Prediction1: edit the icon theme
Prediction2: for all say, folders

Who increases the most, becoming 5.8% more than the test set. This indicates that ESA works better at Who-type questions and improves speaker-related understanding.

Then we perform an error analysis to find how our method helps fix wrong cases of baseline. (i) 26.0% of the samples show obvious discourse relationships between the utterance that best matches the question and the utterance that contains the answer. Most have direct discourse relationships that are annotated while few have two-hop relationships.

(ii) 43.2% of the samples need enhanced attention to speakers. (1) The question is led by the special interrogative word Who, Whose. (2) The question contains prompts of speaker names. (3) Other situations, e.g., the wrong answer from the baseline contains an irrelevant name.

(iii) 9.6% of the incorrect answers from the baseline show strong overconfidence in matching. For example, the answer overlaps the question.

Fig. 8 shows examples of the above situations. Please note that the three phenomena are not mutually exclusive.

Example 1: The question is led by Who, and both the two answers are a speaker name. The answer given by the baseline model is timfrost, which is the nearest speaker to put au on all except for security, referred by the question. But the answer should be lightbright, the speaker of U_1 which contains the mentioned phrase.

Example 2: For this Why-type question, the answer from the baseline is cant save the file, following permission denied. This shows overconfidence in the question-context matching. But our ESA helps to find the gold answer, because it’s owned by root. An Explanation discourse relation between the two utterances is modeled.

Example 3: This wrong case shows a mix of matching confidence, misunderstanding of the speaker transition and the discourse relation. U_0 is an interrogative sentence of qkslwrvolf, which implies that the answer may appear after a speaker transition. The QAP relationship between U_0 and U_1 is another indication of the answer position. However, the baseline answer is a sequence that closely follows get filepath manager to use a different icon.

To sum up, our model solves error cases made by the baseline model, which require a better understanding of speaker transitions and utterance relations. This suggests that ESA enhances speaker-aware features and helps to improve dialogue comprehension.

G. Visualization

In this section, we visualize some intermediate outputs to display the effects of ESA. We use Example 2 in Section V-F for illustration. Fig. 9 shows the representations of utterance before and after Discourse Graph and Speaker Graph.

Fig. 9. Utterance representations before and after Discourse Graph & Speaker Graph.
can see that the second darkest color is around the best-matched phrase and its following sequence.

VI. CONCLUSION

In this work, we study machine reading comprehension on multi-party multi-turn dialogues and propose an Enhanced Speaker-Aware approach to model speaker-related features comprehensively. ESA enhances this essential character respectively in different aspects and granularity and further leverages discourse relationships in dialogue MRC. ESA is evaluated on two multi-party multi-turn dialogue benchmarks, Molweni and FriendsQA. Experimental results show the superiority of our method compared to previous work. In addition, we analyze the contribution of each module by ablation study and details of graph implementation. We also discuss the effect of different forms of speaker-related text. And the dialogue-related pre-training objectives are tried on our task. Examples of cases are presented for intuitive illustration. Our work verifies that speaker roles and speaker-aware interrelations are significant characteristics of dialogue contexts.

REFERENCES

[1] S. Afantenos, E. Kow, N. Asher, and Jérey Perret, “Discourse parsing for multi-party chat dialogues,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 928–937.

[2] S. Afantenos, E. Kow, N. Asher, and Jérey Perret, “Discourse parsing for multi-party chat dialogues,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 928–937.

[3] A. Fred A., “Deep learning using rectified linear units (ReLU),” 2018, arXiv:1803.08375.

[4] N. Asher and A. Lascarides, Logics of Conversation. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[5] N. Asher, J. Hunter, M. Morey, F. Benamara, and S. Afantenos, “Discourse structure and dialogue acts in multiparty dialogue: The stac corpus,” in Proc. 10th Int. Conf. Lang. Resour. Eval., 2016, pp. 2721–2727.

[6] H. Bai and H. Zhao, “Deep enhanced representation for implicit discourse relation recognition,” in Proc. 27th Int. Conf. Comput. Linguistics, 2018, pp. 571–583.

[7] Chloé Braud, M. Convoix, and A. Sędzida, “Cross-lingual RST discourse parsing,” in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, 2017, pp. 292–304.

[8] D. Cai and H. Zhao, “Pair-aware neural sentence modeling for implicit discourse relation classification,” in Proc. Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst., 2017, pp. 458–466.

[9] Y. Cao, W. Bi, M. Fang, S. Shi, and D. Tao, “A model-agnostic data manipulation method for persona-based dialogue generation,” in Proc. 6th Ann. Meeting Assoc. Comput. Linguistics, 2022, pp. 7984–8002.

[10] Henry Y. Chen, E. Zhou, and Jinhao D. Choi, “Robust coreference resolution and entity linking on dialogues: Character identification on TV show transcripts,” in Proc. 21st Conf. Comput. Natural Lang. Learn., 2017, pp. 216–225.

[11] X. Chen, F. Chen, S. Xu, and B. Xu, “A multi domain knowledge enhanced matching network for response selection in retrieval-based dialogue systems,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2022, pp. 6597–6601.

[12] E. Choi et al., “QuAC: Question answering in context,” in Proc. 2018 Conf. Empirical Methods Natural Lang. Process., 2018, pp. 2174–2184.

[13] C. Clark and M. Gardner, “Simple and effective multi-paragraph reading comprehension,” in Proc. 56th Annua. Meeting Assoc. Comput. Linguistics, 2018, pp. 845–855.

[14] C. Clark and M. Gardner, “Simple and effective multi-paragraph reading comprehension,” in Proc. 56th Annua. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers), Jul. 2018, pp. 845–855, doi: 10.18653/v1/P18-1078.

[15] K. Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” in Proc. 8th Int. Conf. Learn. Representations, Addis Ababa, Ethiopia, Apr. 26-30, 2020, [Online]. Available: https://openreview.net/forum?id=r1xMH1BtvB

[16] L. Cui, Y. Wu, S. Liu, Y. Zhang, and M. Zhou, “MuTual: A dataset for multi-turn dialogue reasoning,” in Proc. 58th Annua. Meeting Assoc. Comput. Linguistics, 2020, pp. 1406–1416.

[17] J. Devlin, Ming-Wei Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2019, pp. 4171–4186.

[18] M. Elson and E. Charniak, “Disentangling chat,” Comput. Linguistics, vol. 36, no. 3, pp. 389–409, Sep. 2010.

[19] Y. Fan, P. Li, F. Kong, and Q. Zhu, “A distance-aware multi-task framework for conversational discourse parsing,” in Proc. 29th Int. Conf. Comput. Linguistics, Oct. 2022, pp. 912–921. [Online]. Available: https://aclanthology.org/2022.coling-1.76

[20] Y. Gao et al., “Discern: Discourse-aware entailment reasoning network,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 2439–2449.

[21] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, “Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation,” in Proc. 9th Conf. Empirical Methods Natural Lang. Process. Int. Joint Conf. Natural Lang. Process., 2020, pp. 154–164.

[22] M. Glass et al., “Span selection pre-training for question answering,” in Proc. 58th Annua. Meeting Assoc. Comput. Linguistics, 2020, pp. 2773–2782.

[23] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in Proc. 14th Int. Conf. Artif. Intell. Statist., 2011, pp. 315–323.

[24] Jia-Chen Gu et al., “Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots,” in Proc. 29th ACM Int. Conf. Inf. Knowl. Manage., 2020, pp. 2041–2044.
[73] B. Wang et al., “Strudy: Structured dialogue summarization for dialogue comprehension,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2022, pp. 4949–4958.

[74] W. Wang, C. H. Steven Hoi, and Shafiq R. Joty, “Response selection for multi-party conversations with dynamic topic tracking,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 6581–6591.

[75] Y. Wang, S. Li, and H. Wang, “A two-stage parsing method for text-level discourse analysis,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, 2017, pp. 154–188.

[76] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 38–45.

[77] Chen Henry Wu, Y. Zheng, X. Mao, and M. Huang, “Transferable persona-grounded dialogues via grounded minimal edits,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2021, pp. 2368–2382.

[78] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li, “Sequential match network: A new architecture for multi-turn response selection in retrieval-based chatbots,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers), Jul. 2017, pp. 496–505, doi: 10.18653/v1/P17-1046.

[79] J. Xu, Z. Gan, Y. Cheng, and J. Liu, “Discourse-aware neural extractive text summarization,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 5021–5031.

[80] J. Xu, Z. Lei, H. Wang, Zheng-Yu Niu, H. Wu, and W. Che, “Discovering dialog structure graph for coherent dialog generation,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 1726–1739.

[81] Y. Xu and H. Zhao, “Dialogue-oriented pre-training,” in Findings Assoc. Comput. Linguistics: ACL-IJCNLP, pp. 2663–2673, 2021.

[82] Y. Xu, H. Zhao, and Z. Zhang, “Topic-aware multi-turn dialogue modeling,” in Proc. 35th AAAI Conf. Artif. Intell., 2021, pp. 14176–14184.

[83] A. Yang and S. Li, “SciDTB: Discourse dependency treebank for scientific abstracts,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, 2018, pp. 444–449.

[84] L. Yang et al., “Jart: Intent-aware response ranking with transformers in information-seeking conversation systems,” in Proc. Web Conf., 2020, pp. 2592–2598.

[85] Z. Yang and Jinho D. Choi, “FriendsQA: Open-domain question answering on TV show transcripts,” in Proc. 20th Annu. SIGdial Meeting Discourse Dialogue, 2019, pp. 188–197.

[86] D. Yu, K. Sun, C. Cardie, and D. Yu, “Dialogue-based relation extraction,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4927–4940.

[87] N. Yu, M. Zhang, and G. Fu, “Transition-based neural RST parsing with implicit syntax features,” in Proc. 27th Int. Conf. Comput. Linguistics, 2018, pp. 559–570.

[88] R. Zhang, H. Lee, Lazaros Polymenakos, and Dragomir Radev, “Addresssee and response selection in multi-party conversations with speaker interaction meters,” in Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11937.

[89] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing dialogue agents: I have a dog, do you have pets too ?,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, 2018, pp. 2204–2213.

[90] Z. Zhang and H. Hao, “Advances in multi-turn dialogue comprehension: A survey,” 2021, arXiv:2103.03125.

[91] Z. Zhang, J. Li, P. Zhu, H. Zhao, and G. Liu, “Modeling multi-turn conversation with deep utterance aggregation,” in Proc. 27th Int. Conf. Comput. Linguistics, 2018, pp. 3740–3752.

[92] Z. Zhang, J. Li, and H. Zhao, “Multi-turn dialogue reading comprehension with pivot turns and knowledge,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 1161–1173, 2021.

[93] Z. Zhang, J. Yang, and H. Zhao, “Retrospective reader for machine reading comprehension,” in Proc. 35th AAAI Conf. Artif. Intell., 2021, pp. 14506–14514.

[94] Y. Zheng, R. Zhang, M. Huang, and X. Mao, “A pre-training based personalized dialogue generation model with persona-sparse data,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 9693–9700.

[95] P. Zhong, C. Zhang, H. Wang, Y. Liu, and C. Miao, “Towards persona-based empathetic conversational models,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 6556–6566.

[96] C. Zhu, M. Zeng, and X. Huang, “SDNet: Contextualized attention-based deep network for conversational question answering,” vol. abs/1812.05593, 2018.