Structural Modeling for Dialogue Disentanglement

Xinbei Ma¹,²,³, Zhuosheng Zhang¹,²,³, Hai Zhao¹,²,³,*

¹Department of Computer Science and Engineering, Shanghai Jiao Tong University
²Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
³MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

sjtumaxb@sjtu.edu.cn, zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Tangled multi-party dialogue context leads to challenges for dialogue reading comprehension, where multiple dialogue threads flow simultaneously within the same dialogue history, thus increasing difficulties in understanding a dialogue history for both human and machine. Dialogue disentanglement aims to clarify conversation threads in a multi-party dialogue history, thus reducing the difficulty of comprehending the long disordered dialogue passage. Existing studies commonly focus on utterance encoding with carefully designed feature engineering-based methods but pay inadequate attention to dialogue structure. This work designs a novel model to disentangle multi-party history into threads, by taking dialogue structure features into account. Specifically, based on the fact that dialogues are constructed through successive participation of speakers and interactions between users of interest, we extract clues of speaker property and reference of users to model the structure of a long dialogue record. The novel method is evaluated on the Ubuntu IRC dataset and shows state-of-the-art experimental results in dialogue disentanglement.

1 Introduction

As the boom of social networking services rapidly facilitates communication among people, group chatting is happening all the time, which generates multi-party dialogues with long histories (Lowe et al., 2015; Zhang et al., 2018; Choi et al., 2018; Reddy et al., 2019; Li et al., 2020a). Different from plain texts that are always formally written by the authors, multi-parity dialogues are organized by distributed participants in an random way; thus exhibit disorder and confusion (Kummerfeld et al., 2019; Elsner and Charniak, 2010; Joty et al., 2019; Shen et al., 2006; Jiang et al., 2018, 2021). As is shown in figure 1, the development of a multi-party chatting dialogue has special characters: 1) Random users successively participate in the dialogue and follow certain topics that they are interested in; thus, the threads of those topics in this dialogue are developed. 2) Users reply to former related utterances and mention involved users, thus brings dependencies among utterances. In a word, the behavior of speakers determines the structure of a dialogue history record.

Due to the aforementioned features of dialogue development, there are always multiple ongoing conversation threads developing in a dialogue history simultaneously, which cause troubles for both human and machine to understand dialogue context or further deal with various reading comprehension tasks (Kummerfeld et al., 2019; Elsner and Charniak, 2010; Joty et al., 2019; Shen et al., 2006; Jiang et al., 2018, 2021). Therefore, to some extent, disentangling context or clustering conversation threads can make an effective prerequisite for downstream tasks on dialogues (Elsner and Charniak, 2010; Liu et al., 2021a; Jia et al., 2020), as shown in figure 1.
which contributes to screening concerned parts for further machine reading comprehension (MRC) applications.

Nevertheless, existing works on dialogue disentanglement remain to be improved (Zhu et al., 2020; Yu and Joty, 2020; Li et al., 2020c), which ignore or pay little attention to special features of dialogues and show sub-optimal performance. Earlier works mainly depend on feature engineering (Kummerfeld et al., 2019; Elsner and Charniak, 2010; Yu and Joty, 2020), and use well-constructed handcrafted features to train a naive classifier (Elsner and Charniak, 2010) or linear feed-forward network (Kummerfeld et al., 2019). Recent works mainly based on two strategies: 1) two-step (Mehri and Carenini, 2017; Zhu et al., 2020; Yu and Joty, 2020; Li et al., 2020c; Liu et al., 2021a) and 2) end-to-end (Tan et al., 2019; Liu et al., 2020). In the two-step method, disentanglement task is divided into matching and clustering, which means firstly matching utterance pairs to detect reply-to relations and then dividing utterances into clusters according to the matching score. In the end-to-end strategy, alternatively, for each conversation thread, a dialogue state is modeled, which is mapped with a new utterance and accordingly updated. At the same time, the new utterance is divided into the best-matched thread.

Recently, Pre-trained Language Models (PrLMs) (Devlin et al., 2019; an, 2019; Clark et al., 2020) have brought prosperity to downstream natural language processing tasks by providing contextualized backbones, based on which various works have combined strong PrLMs with dialogue features for gains on performance (Lowe et al., 2015; Li et al., 2020a; Liu et al., 2021b; Jia et al., 2020; Wang et al., 2020). At the same time, domain-adaptive pre-training has been proposed, adapting language models to in-domain tasks better (Wang et al., 2020; Li et al., 2020b; Xu and Zhao, 2021). Works on dialogue disentanglement also manage to make use of PrLMs for performance improvement (Li et al., 2020c; Zhu et al., 2020).

In this work, we design a new model to deal with the long tangled multi-party dialogues leveraging dialogue structure-aware information and aiming at a better solution for the dialogue disentanglement problem so as to contribute to downstream dialogue MRC tasks. The structure of a multi-party dialogue is based on the actions of speakers according to the process of dialogue development. Thus we extract 1) speaker property and 2) reference among users to characterize dependencies of utterances, which is taken into consideration to help with detection of reply-to relationship. Experiments are conducted on DSTC-8 Ubuntu IRC dataset (Kummerfeld et al., 2019), where the model achieves performance progress and makes a new state-of-art model.

2 Background and Related Work

2.1 Pre-trained Language Models

Pre-trained language models (PrLMs) have brought remarkable achievements in a wide range of natural language processing (NLP) tasks. BERT (Devlin et al., 2019) is one of the pioneers that have achieved significant progress in language understanding tasks. It was pre-trained on a large corpus to learn basic language knowledge on the two self-supervised training objectives, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) (Devlin et al., 2019). Variants of BERT such as RoBERTa (an, 2019) and ELECTRA (Clark et al., 2020) are proposed with stronger capacity. Devoted to NLP tasks, PrLMs often work as a contextualized encoder with some task-oriented layers added on the top. For dialogue disentanglement on multi-party dialogues, previous works concatenate utterances as an input to feed into subsequent layers for detecting relationships of utterances (Zhu et al., 2020; Li et al., 2020c).

2.2 Dialogue-related Machine Reading Comprehension

Dialogue-related machine reading comprehension brings challenges on handling the complicated scenarios from multiple speakers and criss-crossed dependency among utterances (Lowe et al., 2015; Yang and Choi, 2019; Sun et al., 2019; Li et al., 2020a). A dialogue is developed by all involved speakers in a distributed way, where an individual speaker focuses and declares oneself on one of the topics discussed in the conversation, or reply to utterances from other speakers. Therefore, consistency and continuity are broken by tangled reply-to dependencies between non-adjacent utterances (Li et al., 2020a; Jia et al., 2020; Ma et al., 2021; Li et al., 2021), leading to a graph structure which is quite different from the smooth presentation in plain texts. Recently, numbers of works of dialogue-related MRC managed to enhance dialogue structure-aware features to improve the adap-
Chapter Title


tation of language models on dialogue passages (Jia et al., 2020; Ouyang et al., 2021; Ma et al., 2021; Li et al., 2021) and achieve progress compared to methods suitable for plain texts. A wide range of dialogue-related MRC tasks such as response selection (Gu et al., 2020; Liu et al., 2021b), question answering (Ma et al., 2021; Li et al., 2021) and emotion detection (Hu et al., 2021) have been inspired by paying attention to dialogue structure-aware information.

2.3 Dialogue Disentanglement

Dialogue disentanglement (Elsner and Charniak, 2010), which is also referred as conversation management (Traum et al., 2004), thread detection (Shen et al., 2006) or thread extraction (Adams, 2008), has been studied for decades due to the importance for both understanding long multi-party dialogues and assisting in down stream NLP tasks on dialogues.

With the goal of disentangling multi-party dialogue history, a number of methods have been proposed for years. Early works are based on feature encode and clustering algorithms. Well-designed handcraft features are constructed to train simple networks to predict whether a pair of utterances are alike or different, and clustering methods are borrowed for partitioning (Elsner and Charniak, 2010; Jiang et al., 2018). Researches are facilitated by a large-scale, high-quality dataset, Ubuntu IRC, published by Kummerfeld et al. (2019). And then, FeedForward network and pointer network (Vinayals et al., 2015) are used, leading to significant progress. But the improvement still partially relies on handcraft-related features (Kummerfeld et al., 2019; Yu and Joty, 2020). Then the end-to-end strategy is proposed and fills the gap between the two steps (Liu et al., 2020), where dialogue disentanglement problem is modeled as a dialogue state transition process, and utterances are clustered by mapping with the states of each dialogue thread.

Inspired by achievements of pre-trained language models (Devlin et al., 2019; Clark et al., 2020; an, 2019), approaches based on PrLMs are recently proposed (Zhu et al., 2020; Gu et al., 2020), where PrLMs produce contextually encode utterances as backbones and the encoded representations of tokens are used for higher-level operations.

However, attention paid to special features of dialogues is inadequate. Feature engineering-based works represent properties of individual utterances such as time, speakers and topics with naive handcraft methods, and ignore dialogue context (Elsner and Charniak, 2010; Kummerfeld et al., 2019). Masked Hierarchical Transformer (Zhu et al., 2020) utilizes the golden conversation structure to operate attentions on related utterances when training models, which results in exposure bias. DialBERT Li et al. (2020c) is a BERT (Devlin et al., 2019) added a LSTM (Hochreiter and Schmidhuber, 1997) on the top for modeling contextual clues and claim a state-of-art performance. In this work, we propose a new design of method considering structure-aware clues, based on the fact that dialogues are developed according to the behave of speakers, so as to disentangle a multi-party chatting dialogue history. We model dialogues with attention paid to two aspects: 1) speaker properties of each utterances, which helps with the understanding of utterances, and 2) interactions of speakers between utterances, which helps with the development of conversation threads. We evaluated the model on Ubuntu IRC dataset (Kummerfeld et al., 2019) and obtained a state-of-art performance.

3 Methodology

The definition of the dialogue disentanglement task and details of our model are sequentially presented in this section, indicating how we make efforts to deal with the disentanglement task with dialogue structure-aware features.

3.1 Task Formulation

Suppose that we perform disentanglement to a long multi-party dialogue history \( D = \{U_0, U_2, \ldots, U_n\} \), where \( D \) is composed of \( n \) utterances. An utterance includes an identity of speaker and a message sent by this user, thus denoted as \( U_i = \{s_i, m_i\} \). As several threads are flowing simultaneously within \( D \), we define a set of threads \( T = \{t_0, t_2, \ldots, t_n\} \) as a partition of \( D \), where \( t_i = \{U_{0i}, \ldots, U_{ki}\}, 0 \leq i_0 \leq i_1 \leq \ldots \leq i_k \leq n \), denoting a thread of the conversation. In this task, we aim to disentangle \( D \) into \( T \). As indicated before, a multi-party dialogue is constructed by successive participation of speakers, who often reply to former utterances of interest. Thus, a dialogue passage can be modeled as a graph structure whose vertices denote utterances and edges denote reply-to relationships between utterances. Therefore, we focus on finding a parent node for each utterance through inference of reply-to relationship, so as to
discover edges and then determine the graph of a conversation thread.

3.2 Model Architecture

Figure 2 shows the architecture of the proposed model, which is introduced in detail in this part. The model architecture consists of three modules, including text encoder, structural interaction, and context-aware prediction. The utterances from a dialogue history are encoded with a pre-trained language model whose output is then aggregated to context-level in the encoder. The representation is sequentially fed into the structural modeling module, used for dialogue structural features modeling to characterize contexts with speaker-aware and reference-aware features. Then in the prediction module, the model performs a fusion and calculates the prediction of reply-to relationships.

3.2.1 Encoder

**Pairwise encoding** Following previous works (Zhu et al., 2020; Li et al., 2020c), we utilize a pretrained language model e.g. BERT (Devlin et al., 2019) as an encoder for contextualized representation of tokens. Since chatting records are always long and continuous, it is inappropriate to concatenate the whole context as input. Thus, we concatenate an utterance with each candidate separately at the encoder stage, satisfying contextual information from former history.

Assuming that for an utterance $U_i$, we consider former $C$ utterances (including $U_i$ itself) as candidates for parent node of $U_i$, the input of a PrLM is in the form of $[CLS] \ U_{i-1} \ [SEP] \ U_i \ [SEP]$, where $0 \leq j \leq C - 1$. The output is denoted as $H_0 \in \mathbb{R}^{C \times L \times D}$, where $C$ denotes the window length in which former utterances are considered as candidates of the parent, $L$ denotes the input sequence length in tokens, $D$ denotes the dimension of hidden states of the PrLM. Note that there is a situation where the golden parent utterance is beyond the range of $(U_i, U_{C-1})$. We label a self-loop for $U_i$ in this case, which means $U_i$ is a beginning of a new dialogue thread as it is too far from the parent, making $U_i$ a root of the thread, which makes sense in the real world, because when users enter a chatroom, they intend to check a limited number of recent messages and make replies, instead of scanning the whole chatting record.

**Utterance Aggregation** $H_0$ is pairwise contextualized representations of each pair of the utterance $U_i$ and a candidate $U_{i-j}$, which will be aggregated to context-level representation for further modeling. Due to the next sentence prediction information is modeled into the position of $[CLS]$, we simply reserve the representations of $[CLS]$. After concatenating pairwise representations from all candidates, we denote the pairwise representations as $H_1 \in \mathbb{R}^{C \times D}$, where $C$ denotes the window length and $D$ denotes the dimension of hidden states of the PrLM.

3.2.2 Structural Modeling

**Speaker Property Modeling** With the goal of enhancing speaker property of each utterance, we follow the mask-based Multi-Head Self-Attention (MHSA) mechanism to emphasize correlations between utterances from the same speaker. The mask-based MHSA is formulated as follows:

$$A(Q, K, V, M) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V,$$

$$head_k = A(HW_t^Q, HW_t^K, HW_t^V, M),$$

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

where $A, head_k, Q, K, V, M, N$ denote the attention, head, query, key, value, mask, and the number of heads, $H$ denotes the input matrix, and $W_t^Q, W_t^K, W_t^V, W^O$ are parameters. Operator $[\cdot, \cdot]$ denotes concatenation. At this step, we input the aggregated representation $H_1$ with a speaker-aware mask:

$$M[i, j] = \begin{cases} 0, & s_i = s_j \\ -\infty, & \text{otherwise,} \end{cases}$$

$$H_2 = \text{MHSA}(H_1, M),$$

where $s$ denotes the speaker identity, $M$ denotes masks of speaker property. The output of MHSA, $H_2 \in \mathbb{R}^{L \times D}$, has the same dimension with $H_1$. We simply concatenate $H_1$ and $H_2$ and adjust to the same size using a linear layer, resulting in a final output of this module denoted as $H_2 \in \mathbb{C}^{L \times D}$.

**Reference Dependency Modeling** As discussed above, the relation of references between speakers is the most important and straightforward dependency among utterances, for references indicate interactions between users which is the internal motivation of the development of a dialogue record. To this end, we build a matrix to label the references, which is regarded as an adjacency matrix of a graph representation. In the graph of references, a vertice denotes an utterance and an edge for reference dependence. For example, if $U_i$ has a reference to the speaker of $U_j$, then there is an
edge from $v_i$ to $v_j$. Inspired by the wide application of graph convolutional network (GCN) (Kipf and Welling, 2017), we borrow the relation-modeling method of relational graph convolutional network (r-GCN) (Schlichtkrull et al., 2018; Shi and Huang, 2019) in order to enhance the reference dependencies, which can be denoted as follows:

$$h_{l+1}^{(l)} = \sigma \left( \sum_{\text{r} \in \mathbb{R}} \sum_{j \in N_j^r} \frac{1}{c_{i,r}} W^{(l)}_r h^{(l)}_j + W^{(l)}_0 h^{(l)}_i \right),$$

where $\mathbb{R}$ is the set of relationships, which in our module is only the reference dependency. $N_j^r$ denotes the set of neighbours of vertex $v_i$, which are connected to $v_i$ through relationship $r$, and $c_{i,r}$ is constant for normalization. $W^{(l)}_r$ and $W^{(l)}_0$ are parameter matrix of layer $l$. $\sigma$ is activated function, which in our implementation is ReLU (Glorot et al., 2011; Agarap, 2018). The input of r-GCN is $H_2$, and after this dependency modeling, the representation is denoted as $H_3 \in \mathbb{C}^{L \times D}$.

### 3.2.3 Context-aware Prediction

As claimed above, the separated pairwise way of encoding satisfied some context information, to compensate which, we put a Bi-LSTM (Hochreiter and Schmidhuber, 1997) layer for compensating contextual clues within the whole window of candidate parents. At the same time, the dialogue structure aware representation $H_3$ need to be combined with the original representation of $[\text{CLS}]$ $H_0$ for enhancement.

Considering both of them, we employ a Syn-LSTM proposed by Xu et al. (2021), which was originally proposed for named entity recognition (NER). A Syn-LSTM models the contextual information while the reference dependency is highlighted, enriching relations among parent candidates. Syn-LSTM is distinguished from an additional input gate for an extra source of input, whose parameters are obtained from training, achieving a better fusion of input sources. The process in a Syn-LSTM cell can be formulated as:

$$f_t = \sigma(W^{(f)} x_{1t} + U^{(f)} h_{t-1} + Q^{(f)} x_{2t} + b_f),$$

$$o_t = \sigma(W^{(o)} x_{1t} + U^{(o)} h_{t-1} + Q^{(o)} x_{2t} + b_o),$$

$$i_t = \sigma(W^{(i)} x_{1t} + U^{(i)} h_{t-1} + b_i),$$

$$c_{1t} = \text{tanh}(W^{(c)} x_{1t} + U^{(c)} h_{t-1} + b_c),$$

$$c_{2t} = \text{tanh}(W^{(p)} x_{2t} + U^{(p)} h_{t-1} + b_p),$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_{1t} + c_{2t},$$

$$h_t = o_t \odot \text{tanh}(c_t),$$

where $x_{1t}$ and $x_{2t}$ are inputs, $f_t$ is a forget gate, $o_t$ is an output gate, $i_t$ and $c_{2t}$ are input gates, and $W$ represents parameter. We use Syn-LSTM in the bi-directional way and the output of Syn-LSTM is $H_4 \in \mathbb{R}^{L \times 2D_r}$, where $D_r$ is the hidden size of Syn-LSTM.

At this stage, $H_4$ is the structure feature-enhanced representation of each pair of the ut-
| Model                        | VI | ARI | 1-1 | F1  | P  | R  |
|-----------------------------|----|-----|-----|-----|----|----|
| **Test Set**                |    |     |     |     |    |    |
| FeedForward (Kummerfeld et al., 2019) | 91.3 | -   | 75.6 | 36.2 | 34.6 | 38.0 |
| ×10 union (Kummerfeld et al., 2019) | 86.2 | -   | 62.5 | 33.4 | 40.4 | 28.5 |
| ×10 vote (Kummerfeld et al., 2019) | 91.5 | -   | 76.0 | 38.0 | 36.3 | 39.7 |
| ×10 intersect (Kummerfeld et al., 2019) | 69.3 | -   | 26.6 | 32.1 | 67.0 | 21.1 |
| Elsner (Elsner and Charniak, 2008) | 82.1 | -   | 51.4 | 15.5 | 12.1 | 21.5 |
| Lowe (Lowe et al., 2017)     | 80.6 | -   | 53.7 | 8.9  | 10.8 | 7.6  |
| BERT (Li et al., 2020c)      | 90.8 | 62.9 | 75.0 | 32.5 | 29.3 | 36.6 |
| DialBERT (Li et al., 2020c)  | 92.6 | 69.6 | 78.5 | 42.3 | 42.3 | 46.2 |
| +cov (Li et al., 2020c)      | 93.2 | 72.8 | 79.7 | 44.8 | 21.1 | 47.9 |
| +feature (Li et al., 2020c)  | 92.4 | 66.6 | 77.6 | 42.2 | 38.8 | 46.3 |
| +future context (Li et al., 2020c) | 92.3 | 66.3 | 79.1 | 42.6 | 40.0 | 45.6 |
| Ptr-Net (Yu and Joty, 2020)  | 92.3 | 70.2 | -   | 36.0 | 33.0 | 38.9 |
| + Joint train (Yu and Joty, 2020) | 93.1 | 71.3 | -   | 39.7 | 37.2 | 42.5 |
| + Self-link (Yu and Joty, 2020) | 93.0 | 74.3 | -   | 41.5 | 42.2 | 44.9 |
| + Joint train&Self-link (Yu and Joty, 2020) | 94.2 | 80.1 | -   | 44.5 | 44.9 | 44.2 |
| BERT_{base} (Our baseline)  | 91.4 | 60.8 | 74.4 | 37.2 | 34.0 | 41.2 |
| Our model                   | 94.0±2.6 | 74.4±13.6 | 81.9±7.5 | 46.0±8.9 | 46.1±12.1 | 47.6±6.4 |

| Dev Set                     |    |     |     |     |    |    |
|-----------------------------|----|-----|-----|-----|----|----|
| Decom. Atten. (Parikh et al., 2016) | 70.3 | -   | 39.8 | 0.6  | 0.9 | 0.7  |
| +feature(Parikh et al., 2016)   | 87.4 | -   | 66.6 | 21.1 | 18.2 | 25.2 |
| ESIM (Chen et al., 2017)       | 72.1 | -   | 44.0 | 1.4  | 2.2 | 1.8  |
| +feature (Chen et al., 2017)   | 87.7 | -   | 65.8 | 22.6 | 18.9 | 28.3 |
| MHT (Zhu et al., 2020)         | 82.1 | -   | 59.6 | 8.7  | 12.6 | 10.3 |
| +feature (Zhu et al., 2020)    | 89.8 | -   | 75.4 | 35.8 | 32.7 | 34.2 |
| DialBERT (Li et al., 2020c)    | 94.1 | 81.1 | 85.6 | 48.1 | 49.5 | 46.6 |
| BERT_{base} (Our baseline)    | 91.7 | 74.6 | 80.2 | 33.5 | 32.2 | 35.0 |
| Our model                    | 94.4±2.7 | 81.8±7.2 | 86.1±5.9 | 52.6±19.1 | 51.0±18.8 | 54.3±9.3 |

Table 1: Experimental results on the Ubuntu IRC dataset (Kummerfeld et al., 2019).

terance $U_i$ and a candidate parent utterance $U_{i-j}$. To measure the correlations of these pairs, we follow previous work (Li et al., 2020c) to consider the Siamese architecture of each pair of $[U_i, U_{i-j}]$ $1 \leq j \leq C - 1$ and the pair of $[U_i, U_i]$:

$$H_{5j} = [p_{ii}, p_{ij}, p_{ii} \odot p_{ij}, p_{ii} - p_{ij}],$$

where $p_{ij}$ is the representation for pair of $[U_i, U_{i-j}]$ from $H_4$, and we get $H_5 \in \mathbb{R}^{L \times 8D_v}$. $H_5$ is fed into a classifier to predict a parent utterance from all parent candidates. We use the cross-entropy for model training.

4 Experiments

The proposed model is evaluated on a large-scale multi-party dialogue record dataset Ubuntu IRC (Kummerfeld et al., 2019), which is also used as a dataset of DSTC-8 Track2 Task4. The results show that our model surpasses the baseline significantly and achieves a new state-of-the-art.

4.1 Dataset

Ubuntu IRC (Internet Relay Chat) is the benchmark corpus published by Kummerfeld et al. (2019), which is collected from #Ubuntu and #Linux IRC channels. Reply-to relations between utterances are manually annotated in the form of (parent utterance, son utterance). Ubuntu IRC consists of 77653 messages and is well-annotated, thus it is the largest and most influential dataset of dialogue disentanglement and contributes to related researches heavily.
4.2 Metrics

**Reply-to relations** We calculated the accuracy for the prediction of parent utterance, indicating the inference ability of reply-to relations.

**Disentanglement** With the goal of dialogue disentanglement, threads of a conversation is formed by clustering all related utterances bridged by reply-to relations, in other words, a connected subgraph. At this stage, we use metrics to evaluate following DSTC-8, which are Variation of Information (VI) \(Kummerfeld \ et \ al., \ 2019\), Adjusted rand index (ARI) \(Kim \ et \ al., \ 2019\), One-to-One Overlap (1-1) \(Elsner \ and \ Charniak, \ 2010\), precision (P), recall (R), and F1 score of clustering. Note that in the table of results, we present 1-VI instead of VI \(Kummerfeld \ et \ al., \ 2019\), thus we expect a larger numerical value for all metrics indicating a stronger performance.

4.3 Setup

Our implementations are based on Transformers Library \(Wolf \ et \ al., \ 2020\). We fine-tune our model employing AdamW \(Loshchilov \ and \ Hutter, \ 2019\) as the optimizer. The learning rate begins with 4e-6. In addition, the input sequence length is set to 128, which our inputs are truncated or padded to, and the window width of considered candidates is 50.

4.4 Experimental Results

Table 1 shows the results of our experiments. The experimental results show that our model outperforms all baselines and previously proposed models by a large margin as highlighted in the table, and achieves a new state-of-the-art (SOTA).

5 Analysis

5.1 Ablation Study

We study the effect of speaker property and reference dependency respectively to verify their contribution. We ablate each of the features and train the model. Results in table 2 show that both speaker property and reference dependency are non-trivial.

5.2 Methods of Aggregation

At the stage of aggregation, we head for context-level representations. We consider the effect of different methods of aggregation, i.e., max-pooling and extraction of \([CLS]\) tokens, the models are trained with the same hyper-parameters. Results in table 3 show our aggregation is the best.

5.3 Layers of LSTM

To see the effect of depth of Syn-LSTM, we did experiments on the numbers of layers of Syn-LSTM, also with the same hyper-parameters. According to the results as shown in Table 4, we put an one-layer LSTM for better performance.

6 Conclusion

In this paper, we study disentanglement on long multi-party dialogue records and propose a new model by paying close attention to dialogue structure, i.e., the speaker property and reference dependency. Our model is evaluated on the largest latest benchmark dataset Ubuntu IRC, where experimental results show the advancement of our method compared to previous work and reach a performance of SOTA. In addition, we analyze the contribution of each structure-related feature by ablation study and the effect of the different model architecture. Our work discloses that speaker property and dependency are significant characters of dialogue contexts and deserves studies in multi-turn dialogue modeling.
References

Holland, Adams, Paige. 2008. Conversation thread extraction and topic detection in text-based chat.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). ArXiv preprint, abs/1803.08375.

Yinhan Liu an. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. ArXiv preprint, abs/1907.11692.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentai Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In Proceedings of ACL-08: HLT, pages 834–842, Columbus, Ohio. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. Computational Linguistics. 36(3):389–409.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 315–323. JMLR Workshop and Conference Proceedings.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, pages 2041–2044. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7042–7052, Online. Association for Computational Linguistics.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1911–1920, Online. Association for Computational Linguistics.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to disentangle interleaved conversational threads with a Siamese hierarchical network and similarity ranking. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1812–1822, New Orleans, Louisiana. Association for Computational Linguistics.

Ziyou Jiang, Lin Shi, Celia Chen, Jun Hu, and Qing Wang. 2021. Dialogue disentanglement in software engineering: How far are we? ArXiv preprint, abs/2105.08887.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Gabriel Murray. 2019. Discourse analysis and its applications. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 12–17, Florence, Italy. Association for Computational Linguistics.

Seokhwan Kim, Michel Galley, R. Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, Minlie Huang, Luis A. Lastras, Jonathan K. Kummerfeld, Walter S. Lasecki, Chiiori Hori, Anoop Cherian, Tim K. Marks, Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkarla, and RagHAV Gupta. 2019. The eighth dialog system technology challenge. CoRR, abs/1911.06394.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.
Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020a. Molweni: A challenge multiparty dialogue-based machine reading comprehension dataset with discourse structure. In Proceedings of the 28th International Conference on Computational Linguistics, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020b. Task-specific objectives of pre-trained language models for dialogue adaptation. arXiv preprint arXiv:2009.04984.

Hui Liu, Zhan Shi, and Xiaodan Zhu. 2020. End-to-end transition-based online dialogue disentanglement. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pages 3868–3874. ijcai.org.

Hui Liu, Zhan Shi, and Xiaodan Zhu. 2021a. Unsupervised conversation disentanglement through co-training. ArXiv preprint, abs/2109.03199.

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2021b. Filling the Gap of Utterance-aware and Speaker-aware Representation for Multi-turn Dialogue. In The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. Dialogue & Discourse, 8(1):31–65.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2021. Enhanced speaker-aware multi-party multi-turn dialogue comprehension. ArXiv preprint, abs/2109.04066.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 615–623, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sira Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Dialogue graph modeling for conversational machine reading. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3158–3169, Online. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings, volume 10843 of Lecture Notes in Computer Science, pages 593–607. Springer.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread detection in dynamic text message streams. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 35–42.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 7007–7014. AAAI Press.
Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics, 7:217–231.

Ming Tan, Dakuo Wang, Yupeng Gao, Haoyu Wang, Saloni Potdar, Xiaoxiao Guo, Shiyu Chang, and Mo Yu. 2019. Context-aware conversation thread detection in multi-party chat. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6456–6461, Hong Kong, China. Association for Computational Linguistics.

David R. Traum, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04), Lisbon, Portugal. European Language Resources Association (ELRA).

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 2692–2700.

Weishi Wang, Steven C.H. Hoi, and Shafiq Joty. 2020. Response selection for multi-party conversations with dynamic topic tracking. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6581–6591, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021. Better feature integration for named entity recognition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3457–3469, Online. Association for Computational Linguistics.

Yi Xu and Hai Zhao. 2021. Dialogue-oriented pre-training. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2663–2673, Online. Association for Computational Linguistics.