Who Says What (WSW): A Novel Model for Utterance-Aware Speaker Identification in Text-Based Multi-Party Conversations

Y. H. P. P. Priyadarshana[a], Zilu Liang[b] and Ian Piumarta[c]
Kyoto University of Advanced Science (KUAS), Kyoto, Japan

Keywords: Multi-Party Written Discourse Analysis, Speaker Identification, Natural Language Processing.

Abstract: Multi-party conversation (MPC) analysis is a growing and challenging research area which involves multiple interlocutors and complex discourse structures among multiple utterances. Even though most of the existing methods consider implicit complicated structures in MPC modelling, much work remains to be done for speaker-centric written discourse parsing under MPC analysis. On the other hand, pre-trained language models (PLM) have achieved a significant success in utterance-interlocutor semantic modelling. In this study, we propose *Who Says What (WSW)*, a novel PLM which models who says what in an MPC to understand equipping discourse parsing in deep semantic structures and contextualized representations of utterances and interlocutors. To our knowledge, this is the first attempt to use the relative semantic distance of utterances in MPCs to design self-supervised tasks for MPC utterance structure modelling and MPC utterance semantic modelling. Experiments on four public benchmark datasets show that our model outperforms the existing state-of-the-art MPC understanding baselines by considerable margins and achieves the new state-of-the-art performance in response utterance selection and speaker identification downstream tasks.

1 INTRODUCTION

Written discourse analysis is important to identify social, political, historical, and cultural backgrounds of dialog systems (Hoey, 2001). This helps to achieve valuable insights such as semantic closeness and written discourse structure of the relevance of text-based conversations. Natural language conversational understanding has received an increasing attention due to its potential value for generation and retrieval-based modelling mechanisms in discourse analysis (Wu, S et al., 2011). Human conversational understanding can be identified as two-party conversations (TPCs) (Wu et al., 2016; Zhou et al., 2018) where two interlocutors are engaged in conversations and multi-party conversations (MPCs) (Traum, 2004; Uthus and Aha, 2013) involving more than two interlocutors. Even though most of the existing methods focus on TPC-based understanding, substantial work has been carried out recently for MPC-based understanding (Hu et al., 2019; Gu et al., 2021). Figure 1 shows the graphical informational flow in an MPC. Sequential-based TPCs learn interlocuter embeddings and map with utterances while graphical-based MPCs build implicit relationships between multiple interlocuter speaker embeddings with respective utterances to identify the dynamic informational flow (Hu et al., 2019). A sample MPC is shown in Table 1.

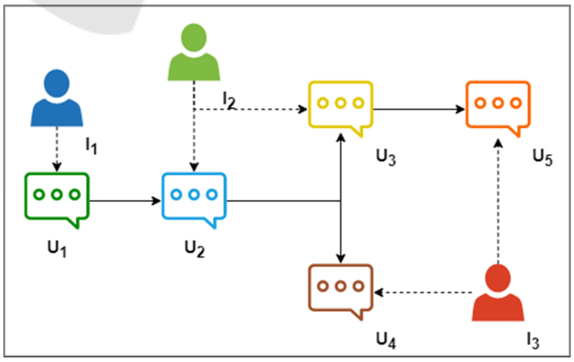

Figure 1: Graphical informational flow of an MPC.

[a] https://orcid.org/0000-0002-4319-3944
[b] https://orcid.org/0000-0002-2328-5016
[c] https://orcid.org/0000-0001-8915-5671

Priyadarshana, Y., Liang, Z. and Piumarta, I.
Who Says What (WSW): A Novel Model for Utterance-Aware Speaker Identification in Text-Based Multi-Party Conversations.
DOI: 10.5220/0012164400003584
In *Proceedings of the 19th International Conference on Web Information Systems and Technologies (WEBIST 2023)*, pages 26-36
ISBN: 978-989-758-672-9; ISSN: 2184-3252

Table 1: Sample of an MPC.

Interlocutors	Utterances
Speaker 1	How are you?
Speaker 2	I'm fine, at least now I'm ok.
Speaker 3	What happened to you?
Speaker 2	I nearly survived, it's shocking.
Speaker 3	What?

The above complicated informational flow between multiple interlocutors (I_1, I_2, and I_3) and multiple utterances (U_1, U_2, U_3, etc.) generates various tasks for MPCs such as identifying the correct speaker of a particular utterance and understanding the discourse structure of the communication flow. Various downstream tasks such as modelling the next speaker, predicting the most appropriate reply, and identifying the addressee of a particular utterance can be identified as core functional units under graphical-based MPCs understanding (Meng et al., 2018; Le et al., 2019). Even though multiple work has been conducted under MPCs analysis to model interlocutor-utterance semantics, a satisfactory study has not been done for speaker-centric discourse parsing under graphical-based MPC analysis (Gu et al., 2022).

Considering the above-mentioned issues, we propose an approach, *WSW*, which models who says what in graphical-based MPCs. Our goal is to design self-supervised tasks on top of pre-trained language models (PLMs) to enhance the PLM's ability to understand MPCs. Here, we introduce a novel approach to equip the discourse parsing considering the relative semantic distance of utterances when designing self-supervised tasks for PLMs. To our knowledge, this is the first attempt to use the relative semantic distance of utterances in MPCs for designing self-supervised tasks for graphical-based MPC understanding. These self-supervised tasks are used to train a Bidirectional Encoder Representations from Transformer (BERT) PLM in a multi-task framework, to model contextual representations of utterances and interlocutors of an MPC. The model is tested on two downstream tasks including *speaker/addressee identification* and *response utterance selection*, to evaluate the generalization and robustness of the model. The main contributions of this study are: (1) A novel PLM on *WSW* is proposed to model contextual representations of utterances and interlocutors of an MPC considering the relative semantic distance of utterances in MPCs. (2) Two downstream tasks are employed to extensively evaluate the generalization and robustness of the proposed model. We will show in the results section that our model outperforms the SOTA MPC understanding models on the above-mentioned downstream tasks with four benchmarks datasets.

The overall organization of the paper is as follows: Section 2 critically reviews related work in this area while Section 3 explains the overall architecture and methodology. Experimental results are presented in Section 4 and then discussed in Section 5.

2 RELATED WORK

Existing approaches on modelling user utterances and interlocutors' identification for MPC understanding can be categorized into MPC-based utterance-aware speaker identification and retrieval-based utterance modelling. Utterance-aware speaker identification identifies the speaker of a specific utterance. Multiple studies have been carried out on speaker identification based on implicit hidden and explicit information of different MPCs utterances. Glass and Bangay introduced a rule-based approach for speaker identification in fictions (Glass and Bangay, 2007). A sequence labelling approach (O'Keefe et al., 2012) and a statistical approach (Elson and McKeown, 2010) were invented for speaker identification of quoted speech in stories to determine who says which line in the extracted text conversations. Meng et al., 2017 proposed the first approach to text-based speaker segmentation considering speaker change detection in a conversation. The speaker change detection is important to identify change of speaker points in a series of utterances. The same approach was modified to segment an MPC based on multiple speakers to classify each speaker for the utterances (Meng et al., 2018). Chen et al., 2021 improved their rule-based approach to identify speakers in novels (Chen et al., 2019), by reformulating it as a BERT-based modelling approach. Although much work has been performed, relative semantic discourse level utterance-aware speaker identification in MPC is yet to be achieved.

The prime objective of retrieval-based utterance modelling is to select the optimal utterance out of all candidates considering the semantic closeness of an MPC. Ouchi and Tsuboi, 2016 proposed a modelling framework for response and addressee selection to identify what has been said to whom in an MPC. Zhang et al., 2018 improved this framework by adding interlocutor embeddings to capture speaker understanding in an MPC but failed to capture the semantic closeness. Wang et al., 2020 were able to achieve significant results in response selection using the concept of dynamic topic tracking without considering the discourse-level semantic distance of each utterance in an MPC. Recently, Gu et al., 2021 introduced a unified multi-task framework for response selection in an MPC but limited to modelling utterances and interlocutor structures without considering the discourse nature of context.

Taking advantage of rapid advancement of PLMs, graphical-based MPCs understanding was further improved. There are studies which proposed integrating topic and interlocutor details into PLMs (Wang et al., 2020) to improve the performance of utterance-interlocutor semantic modelling in MPCs. The utterance-interlocutor semantic modelling is essential to model speaker identification in MPCs for different utterance semantics. Understanding deep semantic structures and contextualized representations of utterances and interlocutors is essential to performing utterance-aware speaker identification in MPCs. Modifying the above-mentioned PLMs considering utterance-interlocutor discourse modelling concepts may produce state-of-the-art (SOTA) results in speaker-aware MPC understanding but is neglected in most studies.

To our knowledge, the present study is the first attempt to use the relative semantic distance of utterances in an MPC to identify *WSW* in a conversation. Relative semantic distance measures the contextual closeness of concepts which is essential in MPCs understanding. Two downstream tasks are used to evaluate the performance of our novel model considering four benchmark datasets.

3 METHODOLOGY

In this section, we present the overall architecture, self-supervised tasks formation for utterance-aware speaker identification along with utterance semantic modelling and downstream tasks formation of our approach.

3.1 Architecture

Figure 2 shows the overall architecture of our model: *WSW.* Although BERT (Devlin et al., 2019) is used as the cornerstone PLM for our model, considering its capability of processing multiple contextualized representations of an MPC, the proposed self-supervised tasks can be modelled with other types of PLMs such as RoBERTa (Liu et al., 2019) for various other fine-tuning purposes.

The structure of an MPC instance consists of three sequential components such as *speaker, utterance,* and *addressee* (Gu et al., 2021). Out of these triplets, our model focusses only on modelling a respective *speaker (S)* with its' *utterance/s (U)*. Once the real MPC data which consist of speaker and utterance details are exposed to the model, those should be encoded to generate input embeddings. These input embeddings can be identified as token embeddings, segment embeddings, speaker embeddings and position embeddings as shown in Figure 2. Token embeddings are essential in understanding utterance level contextual representations while position and segment embeddings are important for feeding the sequential nature and order of input to BERT, since transformers are not aware of the sequential nature as recurrent neural networks (RNN) (Vaswani et al., 2017). Positional and segment-based embeddings are generated randomly considering the inconsistent nature of utterances in an MPC. Speaker embeddings are essential for making BERT aware in identifying related speaker changing information in an MPC (Gu et al., 2020) and are obtained using a context-speaker matching pairs embedding list. This list is updated throughout the model training period considering all utterances of multiple text-based conversations in the dataset.

Once a particular text conversation is picked up during a pre-training cycle, it is processed through two main tasks such as *predicting next utterances* and *predicting masked language inputs*. These are important for obtaining parametric values for satisfying BERT pre-training tasks such as next sentence prediction (NSP) and masked language modelling (MLM) (Devlin et al., 2019). First, a [CLS] token is concatenated at the beginning of each utterance of the chosen conversation to denote the context-level representations and then all utterances are concatenated. The task of *predicting next utterances* is employed to generate the semantic closeness of each concatenated utterance using segment embeddings for comparing two subsequent utterances in the same conversation. A flag called *IS_NEXT* is set to *True* or *False* depending on the obtained semantic closeness value of two subsequent utterances according to Algorithm 1. A [SEP] token is added after processing all concatenated utterances. These tokens are then used for the task of *predicting masked language inputs.* The logic *predicting next utterances* in Algorithm 1 can be identified as the first part of our contribution. Finally, all the obtained response tokens are integrated with context related tokens and then used for self-supervised tasks formation.

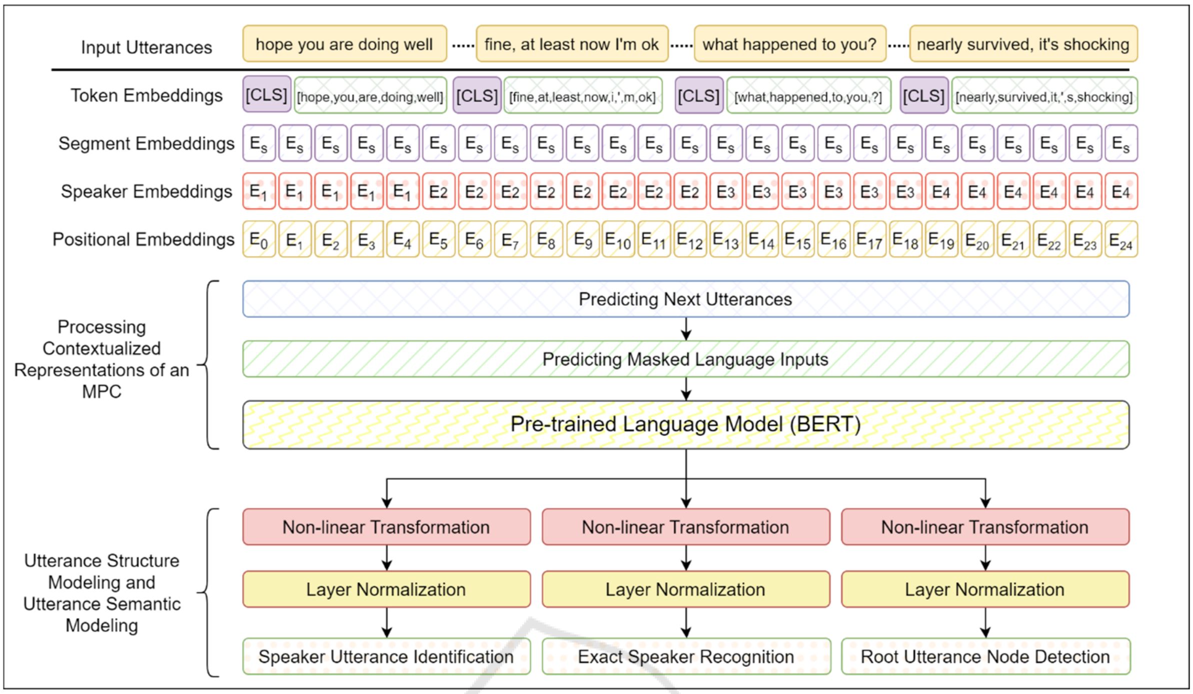
Input Utterances
hope you are doing well
fine, at least now I'm ok
what happened to you?
nearly survived, it's shocking
Token Embeddings
[CLS]
[hope,you,are,doing,well]
[CLS]
[fine,at,least,now,i,',m,ok]
[CLS]
[what,happened,to,you,?]
[CLS]
[nearly,survived,it,',s,shocking]
Segment Embeddings
Es
Speaker Embeddings
E1 E1 E1 E1 E1 E2 E2 E2 E2 E2 E2 E2 E2 E3 E3 E3 E3 E3 E3 E4 E4 E4 E4 E4 E4
Positional Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16 E17 E18 E19 E20 E21 E22 E23 E24
Processing Contextualized Representations of an MPC
Predicting Next Utterances
Predicting Masked Language Inputs
Pre-trained Language Model (BERT)
Utterance Structure Modeling and Utterance Semantic Modeling
Non-linear Transformation
Layer Normalization
Speaker Utterance Identification
Non-linear Transformation
Layer Normalization
Exact Speaker Recognition
Non-linear Transformation
Layer Normalization
Root Utterance Node Detection

Figure 2: Overall model architecture and self-supervised tasks formation of *WSW*.

Algorithm 1: Obtaining semantic closeness of utterances.

```
Semantic Closeness of Utterances
Input: segment_embeddings, utterance_first,
utterance_last
Output: Is_Next
for utterance_first: utterance_last do
  if segment_embeddings[utterance_first] >
segment_embeddings[utterance_last] then
    for i in (utterance_first, utterance_last) do
      if segment_embeddings(i) <
segment_embeddings(i+1) then
        Is_Next <- True;
      else
        Is_Next <- False;
      end if
    end for
  end if
end for
```

3.2 Self-Supervised Tasks Formation

This process can be identified as utterance structure modelling and utterance semantic modelling. Utterance structure modelling consists of two self-supervised tasks: Speaker Utterance Identification *(SUI)* and Exact Speaker Recognition *(ESR)*.

3.2.1 Speaker Utterance Identification (SUI)

This task is proposed to model specific current and preceding utterances of the same speaker in a particular conversation *C*. The encoded tokens for the respective utterances in pre-training cycles are encoded by BERT and are then used to extract the contextualized representations representing those respective utterances. These encoded representations are then processed by applying a non-linear transformation for preserving the linear contextual relationship between current and preceding utterances. A *Dense* layer is then employed to further define relationships of the encoded representations (Ba et al., 2016). A matrix multiplication is performed to generate matching probability scores *(S)* of the contextual representations considering encoded values of a specific current utterance U_i and preceding utterances U_j and can be calculated as follows.

$$S(U)_{ij} = \frac{exp\ (U_i)}{\sum_{j=1}^{n} exp\ (U_j)} \qquad (1)$$

The calculated value defines the matching probability of U_i with its U_j which then is used to minimize the overall loss value of this task. The loss function $\mathcal{L}$ can be determined as

$$\mathcal{L}_{SUI} = -\sum_{i\in C}\sum_{j=1}^{i-1} S'(U)_{ij} log\ (S(U)_{ij}), \qquad (2)$$

Where $S'(U)_{ij}$ denotes the ground truth value for a specific current utterance U_i and preceding utterances U_j.

3.2.2 Exact Speaker Recognition (ESR)

The objective of this task is to predict the speaker of a given utterance using masked speaker embeddings of the pre-training stage.

First, the logic identifies masked speaker positions and masked speaker labels using the positional embeddings of multiple utterances. The positional embeddings represent the relative semantic distance between multiple utterances. After the logic picks utterances that are semantically close, the respective speakers of those selected utterances are determined. The logic in Algorithm 2 can be identified as the second part of our contribution. This novelty addresses the research gap of discourse level speaker identification of each utterance of an MPC.

Algorithm 2: Discourse level speaker identification of an MPC.

```
Discourse Level Speaker Identification
Input: position_embeddings,
speaker_embeddings, context_utterances
Output: masked_speaker_positions,
masked_speaker_labels
masked_speaker_positions <- []
masked_speaker_labels <- []
for i, j in context_utterances do
  masked_utterances[j] <- [i];
  for i in context_utterances do
    if position_embeddings(i) <
position_embeddings(i+1) then
      masked_speaker_positions <-
position_embeddings(i);
      masked_speaker_labels <-
speaker_embeddings(i);
    else
      start <-
position_embeddings[masked_utterances];
      end <-
position_embeddings[masked_utterances + 1];
      for index in (start, end)
        speaker_embeddings(index) <- 0;
      end for
    end if
  end for
end for
```

The identified speaker embeddings are further processed by BERT during model training cycles and then used to extract speaker related contextualized representations. These encoded representations are further processed as in *SUI*, by applying a non-linear transformation and layer normalization. The matching probability scores *(S)* for *ESR* considering a specific current utterance U_i and preceding utterances U_j which share the same speaker, are calculated using the same formula Eq. (1). The overall loss value of this task is determined similarly, Eq. (2).

The next self-supervised task, Root Utterance Node Detection *(RUND)* is designed to accomplish the utterance semantic modelling.

3.2.3 Root Utterance Node Detection (RUND)

MPC utterances can be identified as root-level utterances and sub-level utterances. As illustrated in Figure 1, U_3 and U_4 are sub-level utterances of the U_2 root-level utterance. According to the nature of discourse structure, root-level utterances and sub-level utterances are semantically relevant (Joty et al., 2012). Considering this nature and *SUI*, we design another self-supervised task, *RUND*, to determine the discourse semantic relevance of sub-level utterances.

To accomplish the pre-training process, token level embeddings are prepended considering root-level and sub-level utterances. Once the root-level nodes are identified in multiple utterances, a [CLS] token is concatenated to the beginning of token prepending process while a [SEP] token is used to separate each sub-level utterances. These representations are further encoded by BERT to obtain contextual representation for [CLS]. A SoftMax activation (Sharma et al., 2017) is applied to determine matching probability scores of sub-level utterances representations. Given two sub-level utterances U_i and U_j, the matching probability scores $S(U_{ij})$ are determined similarly to Eq. (1). The loss function $\mathcal{L}$ which is to minimize the overall loss value of *RUND* can be obtained as

$$\mathcal{L}_{RUND} = -[U_{ij}\, log\big(S(U_{ij})\big) + \big(1 - U_{ij}\big) log\,(1 - U_{ij})], \tag{3}$$

where U_{ij} becomes exactly 1 depending on sharing the same root-level utterance between respective sub-level utterances.

3.3 Downstream Tasks Formation

Two downstream tasks such as Reply Utterance Selection *(RUS)* and Speaker Identification *(SI)* are performed based on the three self-supervised tasks.

3.3.1 Reply Utterance Selection (RUS)

This downstream task maps with the self-supervised task of *SUI*. The objective of *RUS* is to determine the semantic relevance of the context and a given reply-to utterance. The context can be modelled as

$$\{(U_n, S_n)\}_{n=1}^{N}, \tag{4}$$

where S_n denotes all speakers and U_n denotes all utterances. The obtained contextual representations from *SUI* are then processed through a non-linear transformation layer applying a dropout layer, and matching probability scores are obtained for the context and the reply-to utterance. A sigmoid activation function (Marreiros et al., 2008) facilitates the entire process. The mean loss is calculated considering the cross-entropy loss of *RUS* relates to the obtained matching probability scores U_{cr} and ground truth labels as follows

$$\mathcal{L}_{RUS} = -[ylog(U_{cr}) + (1-y)log(1-U_{cr})], \tag{5}$$

where y becomes exactly 1 depending on semantic relevance of the context c and a given reply-to utterance r. Finally, accuracy is measured for *RUS* using reported predictions for probability scores and correct predictions for ground truth values.

3.3.2 Speaker Identification (SI)

This downstream task maps with the self-supervised task of *ESR* determining the exact speaker of any given utterance of an MPC. Firstly, extracted positional embeddings are used to gather indexes of contextualized representations of *ESR* using a BERT classifying model. These representations are then processed through a non-liner transformation layer following layer normalization. The matching probability scores U_{Nj} considering masked speakers of all utterances are generated using a *SoftMax* activation function. The loss value can be obtained considering matching probability scores U_{Nj} and ground truth values as

$$\mathcal{L} = -\sum_{j=1}^{N-1} y_{Nj} log(U_{Nj}), \tag{6}$$

where y_{Nj} becomes exactly 1 if both U_j and U_N share the exact same speaker. Finally, accuracy is measured for *SI* using reported predictions for probability scores and correct predictions for ground truth values.

4 EXPERIMENTS

In this section, we present experiments showing the significance of the established self-supervised tasks and downstream tasks of our model.

4.1 Datasets

Our methods are evaluated on four open-access datasets which have been established for benchmarking MPC models. Table 2 presents the sizes of the training, validation, and test sets of the four benchmark datasets used for the experiments.

Table 2: Statistical summary of four benchmark datasets.

Benchmarks		Train	Validation	Test
STAC Corpus[1] (Asher et al., 2016)		30,468	1,000	1,000
ARS Corpus[2] (Ouchi Tsuboi. 2016)	Len-5	461,120	28,570	32,668
	Len-10	495,226	30,974	35,638
	Len-15	489,812	30,815	35,385
GSN Corpus[3] (Hu et al., 2019)		311,725	5,000	5,000
Molweni Corpus[4] (Li et al., 2020)		79,487	4,386	4,430

ARS corpus, GSN corpus and Molweni corpus were constructed based on large-scale Ubuntu multi-party chat conversations (Lowe et al., 2015). STAC corpus is the first dataset which provides a detailed discourse structural representation for an MPC which has been constructed collecting dialogues from online games, and this corpus is essential for evaluating *RUS*. ARS corpus (which is separated into categories such as Len-5, Len-10 and Len-15 based on the session length of a MPC) and GSN corpus are used for evaluating *RUS* and *SI* while Molweni corpus is essential in evaluating discourse structures in MPCs.

4.2 Baselines

The adopted baseline models can be identified as non-pre-trained baseline models and pre-trained baseline models. For the evaluations of *RUS*, we used DRNN

[1] Available at https://www.irit.fr/STAC/corpus.html

[2] Available at https://github.com/hiroki13/response-ranking/tree/master/data/input

[3] Available at https://github.com/morning-dews/GSN-Dialogues

[4] Available at https://github.com/HIT-SCIR/Molweni/tree/main/DP

(Ouchi and Tsuboi, 2016) and SIRNN (Zhang et al., 2018) as non-pre-trained baseline models and BERT (Devlin et al., 2019), RoBERTa (Liu et al. 2019), SA-BERT (Gu et al., 2020) and MPC-BERT (Gu et al., 2021) as pre-trained baseline models. These baselines are recognized as SOTA models for *RUS* understanding in utterance-interlocutor modelling.

For the evaluations of *SI*, we used SMN (Wu et al. 2016), DAM (Zhou et al. 2018), DUA (Zhang et al. 2018) and IoI (Tao et al. 2019) as non-pre-trained baseline models while BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), ELECTRA (Clark et al. 2020), SA-BERT (Gu et al., 2020), MDFN (Liu et al., 2021) and MPC-BERT (Gu et al., 2021) were adopted as pre-trained baseline models. These models have been identified as the SOTA models for identifying the exact speaker of any given utterance of an MPC. Certain models were reconfigured in accordance with the pre-training settings and corpora to make fundamentals equal with each other.

4.3 Implementation Details

Python 3.8 was used as the main programming language and TensorFlow 2.10 (Abadi et al., 2016) was used as the machine learning framework. The GSN corpus (Hu et al., 2019) and Molweni Corpus (Li et al., 2020) were used for pre-training. The maximum sequence length and the maximum utterance number was set to 230 and 7, respectively. The warmup proportion and the learning rate was set to 0.1 and 0.0005, respectively. GELU (Hendrycks and Gimpel, 2016) was used as the activation function for non-linear transformations while Adam methodology (Kingma and Ba, 2014) was used for optimization purposes. BERT was pre-trained for 10 epochs using a GeForce RTX 3090 Ti 24G GPU by setting the batch size to 8. The code is published for replicating the results for further research purposes[5].

4.4 Evaluation Metrics and Results

We can determine respective metrics for evaluating two downstream tasks, Reply Utterance Selection *(RUS)* and Speaker Identification *(SI).*

4.4.1 RUS

Recall was used as the main metric for evaluating *RUS* since $R_n@k$ was adopted in most of the SOTA models (DRNN (Ouchi and Tsuboi, 2016) and SIRNN (Zhang et al., 2018)) for identifying reply-to selection. $R_n@k$ is an enhanced version of recall considering *k* best-matched reply-to utterances out of *n* available candidates. The setup was performed by setting *k* to 1 and *n* to 2 or 10, respectively[6]. This was to set the optimal fundamentals for modelling specific current and preceding utterances of the same speaker in a particular conversation *C*. Table 3 shows the evaluation results of *RUS* compared to the SOTA models. Ablation tests were also performed considering the self-supervised tasks of our proposed model for performance comparison.

4.4.2 SI

Precision@1, recall ($R_n@k$) and, F_1 score were used as the evaluation metrics for evaluating the *SI* task. Precision@1 (P@1) is identified as the highest ranked precision which is the most natural way of evaluating the performance of a task determining the exact speaker of any given utterance of an MPC with respect to ground truth (Le et al., 2019). P@1 was employed for evaluations with SOTA models using GSN Corpus (Hu et al., 2019) and ARS Corpus (Ouchi and Tsuboi. 2016). Table 4 shows the evaluation results of *SI* in terms of P@1. Our model performed well at a significant margin even when the length of a session was increased, and this is a unique achievement in *SI* where other SOTA models performed in the opposite way. Additionally, ablation tests were also performed considering the self-supervised tasks of our proposed model for performance comparison.

$R_n@k$ was adopted as the evaluation metric for the *SI* task using GSN Corpus (Hu et al., 2019) and STAC Corpus (Asher et al., 2016). The setup was performed by setting *k* to 1, 2 and 5 and *n* to 2 or 10, respectively. This was to set the optimal fundamentals for determining the exact speaker of any given utterance of an MPC. Table 5 shows the evaluation results of *SI* in terms of $R_n@k$.

F_1 score was used for *SI* evaluations with SOTA models (SA-BERT (Gu et al., 2020), MDFN (Liu et al., 2021), and Ma et al., 2021) using GSN corpus (Hu et al., 2019) and Molweni Corpus (Li et al., 2020). F_1 score was employed since it was used by most of the previous SOTA models for *SI* recognition in MPC analysis. Table 6 show the evaluation results of *SI* in terms of F_1 score. The results show that our proposed enhancements for *SI* in MPCs outperformed the SOTA models by a significant margin.

[5] https://github.com/CyraxSector/WSW

[6] These settings were used in previous studies.

Table 3: Evaluation results of RUS in terms of R$_n$@k. Non-pre-trained models and pre-trained models are shown in the 1st and 2nd row, respectively while ablation results are shown in the last row.

	GSN Corpus		ARS Corpus					
			Len-5		Len-10		Len-15	
	R$_2$@1	R$_{10}$@1	R$_2$@1	R$_{10}$@1	R$_2$@1	R$_{10}$@1	R$_2$@1	R$_{10}$@1
DRNN (Ouchi and Tsuboi, 2016)	-	-	76.23	33.89	78.72	36.58	79.13	37.21
SIRNN (Zhang et al., 2018a)	-	-	78.52	36.82	80.48	39.67	81.28	41.27
BERT (Devlin et al., 2019)	93.24	73.83	85.83	54.31	87.27	57.82	87.58	59.21
RoBERTa (Liu et al. 2019)	93.28	74.92	86.26	55.57	88.16	58.34	88.37	60.24
SA-BERT (Gu et al., 2020)	93.36	75.62	86.83	55.51	88.21	59.68	88.64	60.73
MPC-BERT (Gu et al., 2021)	94.95	79.13	87.81	58.21	89.52	62.21	89.83	63.83
WSW (Ours)	**96.53**	**84.27**	**89.34**	**79.28**	**91.38**	**78.24**	**92.43**	**80.63**
WSW w/o. *SUI*	96.21	84.17	89.14	78.83	91.16	77.86	91.47	80.24
WSW w/o. *ESR*	96.38	84.21	89.27	79.18	91.25	78.06	92.29	80.51
WSW w/o. *RUND*	95.53	82.49	87.46	78.53	90.83	77.21	90.76	79.62

Table 4: Evaluation results of SI in terms of P@1. Pre-trained models are shown in the 1st row while ablation results are shown in the last row.

	GSN Corpus	ARS Corpus		
		Len-5	Len-10	Len-15
BERT (Devlin et al., 2019)	71.81	62.24	53.17	51.58
RoBERTa (Liu et al. 2019)	73.46	63.85	55.24	53.72
SA-BERT (Gu et al., 2020)	75.88	64.96	57.62	54.28
MPC-BERT (Gu et al., 2021)	83.54	67.56	61.00	58.52
WSW (Ours)	**85.67**	**68.94**	**69.32**	**70.36**
WSW w/o. *SUI*	84.53	68.37	68.58	69.54
WSW w/o. *ESR*	82.37	66.94	65.52	65.39
WSW w/o. *RUND*	85.38	68.62	68.93	69.85

Table 5: Evaluation results of SI in terms of R$_n$@k. Non-pre-trained models and pre-trained models are shown in the 1st and 2nd row, respectively while ablation results are shown in the last row.

	GSN Corpus		STAC Corpus			
	R$_2$@1	R$_{10}$@1	R$_2$@1	R$_{10}$@1	R$_{10}$@2	R$_{10}$@5
SMN (Wu et al. 2017)	-	-	62.62	58.61	61.76	66.24
DAM (Zhou et al. 2018)	-	-	64.51	59.82	63.72	68.28
DUA (Zhang et al. 2018)	-	-	-	59.24	64.83	68.61
IoI (Tao et al. 2019)	-	-	65.53	62.38	64.26	69.34
BERT (Devlin et al., 2019)	81.54	72.43	70.64	65.24	68.31	72.16
RoBERTa (Liu et al. 2019)	82.38	73.29	71.81	64.35	68.24	73.42
ELECTRA (Clark et al. 2020)	-	-	72.38	66.41	69.62	74.81
SA-BERT (Gu et al., 2020)	85.52	76.28	73.54	67.28	70.34	76.11
MDFN (Liu et al., 2021)	-	-	74.25	68.31	70.95	76.61
MPC-BERT (Gu et al., 2021)	87.75	80.83	77.92	69.46	73.62	78.26
WSW (Ours)	**89.67**	**82.67**	**80.24**	**71.67**	**77.69**	**82.38**
WSW w/o. *SUI*	85.68	79.56	77.92	68.61	76.27	81.52
WSW w/o. *ESR*	83.51	76.18	75.33	64.88	73.66	79.61
WSW w/o. *RUND*	86.62	81.74	79.81	70.72	76.82	81.78

Table 6: Evaluation results of speaker identification in terms of F_1 score.

	GSN Corpus	Molweni Corpus
BERT (2019)	79.54	60.62
RoBERTa (2019)	81.62	60.92
SA-BERT (2020)	82.46	62.48
MDFN (2021)	84.67	64.27
Ma et al. (2021)	85.31	65.83
WSW (Ours)	**87.56**	**69.47**

5 DISCUSSION

In this work, we present a novel PLM called *WSW* to model contextual representations of utterances and interlocutors of an MPC considering the relative semantic distance of utterances. To our knowledge, this is the first significant work which was done for modelling speaker-centric discourse parsing under graphical-based MPC analysis. Three self-supervised tasks were designed for MPC utterance structure modelling and MPC utterance semantic modelling. Two downstream tasks were employed to extensively evaluate the generalization and robustness of the novel PLM. Ablation results on *RUS* and *SI* showed that each self-supervised task is essentially necessary for the model performance. Four public benchmark datasets were used for SOTA model evaluations.

Experiments on *RUS* showed that our model performed better by a significant margin compared to the SOTA PLM and non-PLM models. Table 3 shows that *WSW* outperformed the existing top-performer, MPC-BERT by considerable margins of 5.64%, 21.07%, 16.03% and 16.8% in terms of $R_{10}@1$. Our PLM performed even better when the length of a session was increased which can be identified as a significant observation in overall MPC analysis. The reason for this observation is that our *WSW* considers the discourse structure of utterance-interlocutor modelling which was neglected in other SOTA models. Ablation results under *RUS* evaluations showed that *RUND* self- supervised task contributed more to reply-to utterance selection in MPC. In other words, removing *SUI* or *ESR* from the main logic would not make a significant impact on the performance of *RUS* modelling.

Experiments on *SI* were conducted employing Precision@1, Recall ($R_n@k$), and F_1 score as experimental metrics. According to Table 4, our model outperformed MPC-BERT by considerable margins of 2.13%, 1.38%, 8.32% and 11.84% in terms of P@1. Again, our PLM performed even better when the length of a session was increased which can be identified as a significant observation in overall speaker modelling in MPC analysis. Ablation results under *SI* evaluations showed that the *ESR* self-supervised task contributed more to speaker identification in MPC. According to Table 5, our model outperformed MPC-BERT by margins of 1.84% and 2.21% in terms of $R_{10}@1$. Ablation results confirmed that *ESR* self- supervised task contributed more to speaker identification in MPC. Evaluation results in Table 6 show that *WSW* performed significantly well in F_1 score by margins of 2.25% and 3.64%, respectively, in the GSN corpus (Hu et al., 2019) and Molweni Corpus (Li et al., 2020). In summary, we can conclude that our novel PLM outperformed the existing SOTA models for speaker-interlocutor modelling in MPC analysis.

Although our model focuses only on modelling a respective *speaker (S)* with its' *utterance/s (U),* few other MPC understanding SOTA solutions modelled all components of an MPC such as *speaker, utterance,* and *addressee.* This can be identified as a limitation as well as a potential future enhancement. Our novelties in utterance-interlocutor discourse modelling using relative semantic distance of utterances can be further enhanced to address all components of an MPC. Considering the dynamics of MPC analysis, the minimum number of participants and the maximum number of participants of a given conversation was set to 3 and 7, respectively. Further experiments can be conducted for possible different combinations of minimum/maximum participants while enhancing the overall MPC structure and semantic modelling.

The universal MPC understanding can be identified as a future direction for designing better self-supervised tasks considering the MPC discourse parsing. This will eventually lead to constructing more specialized downstream tasks such as opinion mining of the speakers of an MPC structure. Adopting prompt-learning can be identified as another potential direction where sentence-level prompt-learning can be used to enhance the contextualized MPC NSP logic in PLMs (Ding et al., 2021). Another future enhancement will be applying MPC modelling to low- or zero-resource modelling, which has not been investigated much to date (Gu et al., 2022).

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016, November). Tensorflow: a system for large-scale machine learning. In *Osdi* (Vol. 16, No. 2016, pp. 265-283).

Asher, N., Hunter, J., Morey, M., Benamara, F., & Afantenos, S. (2016, May). Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2721-2727).

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450.*

Chen, J. X., Ling, Z. H., & Dai, L. R. (2019). A Chinese Dataset for Identifying Speakers in Novels. In *INTERSPEECH* (pp. 1561-1565).

Chen, Y., Ling, Z. H., & Liu, Q. F. (2021). A Neural-Network-Based Approach to Identifying Speakers in Novels. In *Interspeech* (pp. 4114-4118).

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H. T., & Sun, M. (2021). Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998.*

Elson, D., & McKeown, K. (2010, July). Automatic attribution of quoted speech in literary narrative. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 24, No. 1, pp. 1013-1019).

Glass, K., & Bangay, S. (2007, November). A naive salience-based method for speaker identification in fiction books. In *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa* (PRASA'07) (pp. 1-6).

Gu, J. C., Li, T., Liu, Q., Ling, Z. H., Su, Z., Wei, S., & Zhu, X. (2020, October). Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 2041-2044).

Gu, J. C., Tao, C., Ling, Z. H., Xu, C., Geng, X., & Jiang, D. (2021). MPC-BERT: A pre-trained language model for multi-party conversation understanding. *arXiv preprint arXiv:2106.01541.*

Gu, J. C., Tao, C., & Ling, Z. H. (2022). Who says what to whom: A survey of multi-party conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* (IJCAI-22).

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136.*

Hoey, M. (2001). Textual interaction: An introduction to written discourse analysis. Psychology Press.

Hu, W., Chan, Z., Liu, B., Zhao, D., Ma, J., & Yan, R. (2019). Gsn: A graph-structured network for multi-party dialogues. *arXiv preprint arXiv:1905.13637.*

Joty, S., Carenini, G., & Ng, R. (2012, July). A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 904-915).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Le, R., Hu, W., Shang, M., You, Z., Bing, L., Zhao, D., & Yan, R. (2019, November). Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. 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) (pp. 1909-1919).

Li, J., Liu, M., Kan, M. Y., Zheng, Z., Wang, Z., Lei, W., ... & Qin, B. (2020). Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Liu, L., Zhang, Z., Zhao, H., Zhou, X., & Zhou, X. (2021, May). Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 15, pp. 13406-13414).

Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909.*

Ma, X., Zhang, Z., & Zhao, H. (2021). Enhanced speaker-aware multi-party multi-turn dialogue comprehension. *arXiv preprint arXiv:2109.04066.*

Marreiros, A. C., Daunizeau, J., Kiebel, S. J., & Friston, K. J. (2008). Population dynamics: variance and the sigmoid activation function. *Neuroimage, 42*(1), 147-157.

Meng, Z., Mou, L., & Jin, Z. (2017, November). Hierarchical RNN with static sentence-level attention for text-based speaker change detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2203-2206).

Meng, Z., Mou, L., & Jin, Z. (2018, April). Towards neural speaker modeling in multi-party conversation: The task, dataset, and models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., & Honnibal, M. (2012, July). A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in*

Natural Language Processing and Computational Natural Language Learning (pp. 790-799).

Ouchi, H., & Tsuboi, Y. (2016, November). Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2133-2143).

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci,* 6(12), 310-316.

Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., & Yan, R. (2019, July). One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1-11).

Traum, D. (2004). Issues in multiparty dialogues. In *Advances in Agent Communication: International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003. Revised and Invited Papers* (pp. 201-211). Springer Berlin Heidelberg.

Uthus, D. C., & Aha, D. W. (2013). Multiparticipant chat analysis: A survey. *Artificial Intelligence, 199*, 106-121.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Wang, W., Joty, S., & Hoi, S. C. (2020). Response selection for multi-party conversations with dynamic topic tracking. *arXiv preprint arXiv:2010.07785*.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, March). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 705-714).

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2016). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Zhang, R., Lee, H., Polymenakos, L., & Radev, D. (2018, April). Addressee and response selection in multi-party conversations with speaker interaction rnns. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W. X., ... & Wu, H. (2018, July). Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1118-1127).