DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement

Tianda Li[1]*, **Jia-Chen Gu**[2]*, **Xiaodan Zhu**[1], **Quan Liu**[2,3], **Zhen-Hua Ling**[2], **Zhiming Su**[3], **Si Wei**[3]
[1]ECE & Ingenuity Labs, Queen's University, Kingston, Canada
[2]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China
[3] State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Hefei, China
{tianda.li,xiaodan.zhu}@queensu.ca, gujc@mail.ustc.edu.cn,
{quanliu,zhling}@ustc.edu.cn, {zmsu,siwei}@iflytek.com

Abstract

Disentanglement is a problem in which multiple conversations occur in the same channel simultaneously, and the listener should decide which utterance is part of the conversation he will respond to. We propose a new model, named Dialogue BERT (DialBERT), which integrates local and global semantics in a single stream of messages to disentangle the conversations that mixed together. We employ BERT to capture the matching information in each utterance pair at the utterance-level, and use a BiLSTM to aggregate and incorporate the context-level information. With only a 3% increase in parameters, a 12% improvement has been attained in comparison to BERT, based on the F1-Score. The model achieves a state-of-the-art result on the a new dataset proposed by IBM (Kummerfeld et al., 2019) and surpasses previous work by a substantial margin.

1 Introduction

With the growth of the Internet, single stream conversation occurs everyday. But in many cases, all messages are entangled with each other, which makes it difficult for a new user to the chat in understanding the context of discussion in the chat room. Automatic disentanglement can help people leverage this type of problem. Research relevant to disentanglement has been conducted over a decade. Aoki et al. (2003, 2006) attempted to disentangle conversation speech. Then, several studies were conducted in text-based conversational media. The primary mainstream strategy of solving disentanglement is modeled as the Topic Detection and Tracking (TDT) task (Allan, 2002), which calculates sentence pair similarity iteratively and decides which conversation the message belongs to. In addition to neural networks, statistical (Du et al., 2017) and linguistic features (Elsner and Charniak, 2008, 2010, 2011; Mayfield et al., 2012) are also used in existing research. Without considering relationships between words in a message, the quality of features will affect the final result dramatically. Most previous works either used small dataset (Elsner and Charniak, 2008) or used an unpublished dataset (Adams and Martell, 2008; Jiang et al., 2018).

The publication of a new large-scale dataset (Kummerfeld et al., 2019) made it possible to train a more complex model. Zhu et al. (2019) proposed masked hierarchical transformer based on BERT to learn a better conversation structure. But their approach need conversation structure in advance to formulate the Mask for their transformer model, which is not possible during the test process.

Unsupervised pre-training (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Sanh et al., 2019; Lan et al., 2019) has recently been shown to be very effective in improving the performances of a wide range of NLP tasks (Li et al., 2019). Previous work has also explored the possibility of post-training BERT in the Ubuntu response selection task (Whang et al., 2019).

In this paper, we propose our own pre-trained model, named Dialogue BERT (DialBERT). Conversation disentanglement leverages DialBERT in three steps: (1) Using BERT to capture the matching information in each sentence pair. (2) Using a context-level BiLSTM to incorporate the context information. (3) Measuring the semantic similarities between sentence pairs. Selecting the pair with the highest score and classifying this pair into the same conversation. The usage of context-level BiLSTM enables BERT to capture semantics across a wide range of contexts. With only a 3 % increase in parameters, DialBERT shows an improvement on all evaluation metrics including an improvement of 12% on the F1-score in comparison to BERT

*Equal contribution.

(Devlin et al., 2018). Also, our model surpasses previous baseline models according to all the evaluation metrics by a wide margin, becoming a new state-of-the-art model.

2 Problem Statement

In this section, we formally define the object of this work and notations used.

Every message can be defined as a tuple $m = (u,\ s,\ t)$, where $u = \langle u_1, u_2, ..., u_n \rangle$ is the word sequence posted by speaker $s \in S$ at time t, where S is the set of speakers. Each message belongs to a specific conversation A. Messages from different conversations could occur in the same channel concurrently. Every message has a preceding message, which is an indication of the message being the response to its respective preceding message.

Following the setting of previous works (Elsner and Charniak, 2008, 2010, 2011; Mayfield et al., 2012; Jiang et al., 2018). Our objective for model training is to calculate the similarity between two messages and to identify which message precedes the current message based on the calculated similarity. As Jiang et al. (2018) pointed out, it is not required to calculate all message pair similarities to find out the preceding message. In our work, $T-1$ messages occurring before this target message are taken in to consideration. Target message will be represented by $M = \{m_1, m_2, ..., m_n\}$ $(M \in \mathbb{R}^n)$, and context messages will be represented by $C_t = \{c_1^t, c_2^t, ..., c_m^t\}$ $(C \in \mathbb{R}^{T\times m})$, where n and m are the sequence length of messages and t is the index of the preceding message. $Y \in \{0, 1\}$ indicates if the message is the preceding message of the target message. Our goal is to learn a disentanglement model, which could predict which message in the context range is the preceding message of the target message by minimizing the loss from a given dataset D.

3 Methodology

In this section, we propose our pre-trained DialBERT model to address the disentanglement problem.

3.1 Domain adaptation

We use the Ubuntu forum data to post-train BERT (Devlin et al., 2018), in order to further improve our model performance (details shown in Appendix A.2).

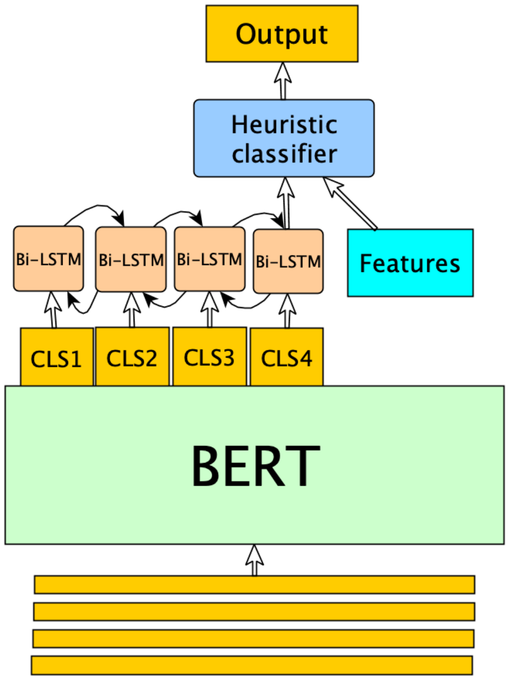

Figure 1: The overall architecture of our proposed model.

To optimize the model, both Masked Language Model likelihood (MLM) and Next Sentence Prediction (NSP) are used as the post-train tasks, and the final loss function is formulated as

$$Loss_{post-training} = Loss_{MLM} + Loss_{NSP}. \quad (1)$$

$$L = -\frac{1}{N}(\sum_{j\in N} y_j logp_j + \alpha(\sum_{i\in N} c_i logp_i)) \quad (2)$$

3.2 DialBERT

We propose DialBERT to calculate the similarity between the target message and context messages.

3.2.1 Context Aware Input

To determine if two messages belong to the same conversation, context should be considered. For every message, we will take $T-1$ messages occurring before the target message in the same channel as its context, and combine the target message to each of its context as input. The similarity score is calculated between the target message and every context message. We concatenate them as follows:

$$Input_i = \left[cls, m_1, m_2.., sep, c_1^i, c_1^i, .., sep\right], \quad (3)$$

where $i \in [0, 1, ..., T-1]$ is the index of the context message. cls and sep are the start and separation tokens predefined in BERT. Note that C_0 is also the target message.

3.2.2 Context BERT module

BERT can encode a sentence with length less than 512 sub-words, which restricts the application of the model when adapted to downstream tasks. We propose a new model called Dialogue BERT(DialBERT) shown in Figure1, which is capable of disentangling a conversation. Every time we package T message pairs(target message with context messages) as one sample. After being encoded by BERT, we use a single layer Bi-LSTM to capture the information given by the encoded output, which means context information will be considered when DialBERT calculates the similarity of message pairs. In addition to BERT, RNN-based Bi-LSTM can help capture semantics across different message pairs. The output f is formulated as:

$$e_i = \textbf{BERT}(Input_i), \quad (4)$$

$$f_i = \textbf{BiLSTM}(e_i), \quad (5)$$

$$\forall i \in [0, 1, ..., T-1], \quad (6)$$

where $f_i \in \mathbb{R}^H$ is the similarity feature vector between the target message and context message with index i. Note that f_0 would be the similarity feature vector calculated between two target messages, which will be used as a "pivot" for classification.

3.2.3 Heuristic Classifier

In order to model higher-order interaction between the target message and context messages, we also compute the difference and element-wise product for the target similarity feature vector (f_0) and every corresponding context similarity feature vector (f_i). We duplicate target similarity feature vectors to form a feature matrix($F^t \in \mathbb{R}^{T\times H}$) having the shape of the context similarity feature matrix($F^c \in \mathbb{R}^{T\times H}$). Along this direction, we also further model this interaction by feeding tuple G into feedforward neural networks with $Tanh$ activation function.

$$F^t = [f_0, f_0, f_0, .., f_0], \quad (7)$$

$$F^c = [f_0, f_1, f_2, .., f_{T-1}], \quad (8)$$

$$G = [F^t, F^c, F^t \cdot F^c, F^t - F^c], \quad (9)$$

$$K = \textbf{Tanh}(\mathbf{G} \times \mathbf{W}^\top + \mathbf{b}), \quad (10)$$

$$P = \textbf{Softmax}(K). \quad (11)$$

	VI	ARI	1-1	F1	P	R
Linear	88.9	-	69.5	21.8	19.3	24.9
Feedforward	91.3	-	75.6	36.2	34.6	38.0
× 10 union	86.2	-	62.5	33.4	40.4	28.5
× 10 vote	**91.5**	-	**76.0**	**38.0**	36.3	**39.7**
× 10 intersect	69.3	-	26.6	32.1	**67.0**	21.1
Elsner(2008)	82.1	-	51.4	15.5	12.1	21.5
Lowe(2017)	80.6	-	53.7	8.9	10.8	7.6
Zhu(2019) [1]	82.1	-	59.6	8.7	12.6	10.3
+feature [1]	89.8	-	75.4	35.8	32.7	34.2
BERT	90.8	62.9	75.0	32.5	29.3	36.6
+ feature	92.2	65.9	76.8	37.8	33.9	42.5
DialBERT	**92.6**	**69.6**	78.5	**44.1**	**42.3**	46.2
+ feature	92.4	64.6	77.6	42.2	38.8	**46.3**
+ future	92.3	66.3	**79.1**	42.6	40.0	45.6

[1] Their models' results are reported according to dev set.

Table 1: Results on Ubuntu test set, our work substantially outperforms prior work by all evaluation metrics. Six types of metrics are considered including the modified Variation of Information (VI) in (Kummerfeld et al. 2019), Adjusted rand index(ARI), One-to-One Overlap (1-1) of the cluster (Elsner and Charniak 2008), and the precision, recall, and F1 score between the cluster prediction and ground truth. Note that precision, recall, and F1 score are calculated using the number of perfectly matching conversations, excluding conversations with only one message (mostly system messages).

$P \in \mathbb{R}^T$ is the similarity score between target message and each context message. We adopt cross-entropy loss for DialBERT.

4 Experimental Setup

In this section, we will introduce the dataset we used to test our model and the setting of our model.

4.1 Data

In 2019, a conversation disentanglement dataset was published. This dataset contains 77563 messages of IRC manually annotated with reply-to relations between messages, which is proposed by (Kummerfeld et al., 2019)(details are shown in A.3). Roughly 75K messages were used, which is still substantially higher than previous work.

4.2 Model

In order to compare our work we employed several baseline models. Following the setting of (Kummerfeld et al., 2019), we adopt linear and feedforward models as baselines. Aiming to import a strong baseline result, union, vote and intersect ensemble strategies given by (Kummerfeld et al., 2019) were used with feedforward model. To make our work more comparable,we also run the experiment using BERT (training details are shown in A.4). Moreover, we also including previous BRET based model's result proposed by (Zhu et al., 2019). An extensive list of the modified pre-trained models are listed below.

- **BERT** BERT model will be used to rank potential antecedents, and find preceding messages according to the ranking scores.
- **BERT + feature** The same setting as BERT, combined with the same features used in Linear model(Kummerfeld et al., 2019). Specifically, The features consist of three part: (1) Global-level features including year and frequency of the conversation. (2) Utterance level features including types of message, targeted or not, time difference between the last message, etc., (3) utterance pair features including how far apart in position and time between the messages, whether one message targets another, etc.
- **DialBERT** DialBERT with adaptation will be used to find preceding message according to the ranking scores.
- **DialBERT + feature** Same linguistic features will be used as **BERT+feature**.
- **DialBERT + future context** In this setting, Not only $T-1$ messages occurred before this target message will be considered, K messages occurred after target message will be considered as well during ranking potential antecedents.

5 Result & Analysis

5.1 Overall Performance

Results are shown in Table 1 [1]. For previous work, an ensemble of 10 feedforward models obtained through a vote were capable of reaching the best performance in 4 out of 5 evaluation metrics.

Different from other NLP tasks, BERT does not perform well, which indicates semantic knowledge learned from pre-training is not a direct indicator of improvement for disentanglement. Post-combination with linguistic features, BERT does have advantages compared with the feedforward neural network according to evaluation metrics, which indicates semantic knowledge acquired by pre-training assists a model in performing better in downstream tasks through proper usage. Another reason might be that linguistic features import information across different sentences, which in turn assist BERT in making better decisions.

[1] Our result is averaged over four runs

	VI	ARI	F1	P	R
Model-AVG	**92.8**	**72.1**	43.3	39.8	47.6
Probability-AVG	92.6	66.6	**45.3**	**42.1**	**49.0**
Vote-AVG	92.6	66.6	45.0	42.0	48.5

Table 2: Results for ensemble models on test set.

The result that DialBERT outperforms BERT on all 6 evaluation metrics could be explained by the vital importance of context in disentangle conversations, and DialBERT makes better use of pre-trained knowledge. For DialBERT, external linguistic features only led to a slight improvement on a single metric, wheres the performance did not improve across other metrics. The reason might be DialBERT is powerful enough to capture most of the information given by those features. Improvements are not shown by importing future context, which maybe attribute to the information acquired by precedent context being sufficient for DialBERT. Dialogue BERT outperforms all previous models on all six evaluation metrics.

In order to find out how well our model could handle the disentanglement problem, We also propose several DialBERT ensembles, the results is given on table 2. Three strategies were put into use, probability-average (Prob-AVG) , vote-average (Vote-AVG), model weight-average (Model-AVG) are put into use (details shows in A.5). The result shows that Prob-AVG and Model-AVG strategies could reach better performance in comparison to vote-average. Model-AVG performs better in 1-Scaled VI and ARI metrics, Prob-AVG performs better in F1, recall and precision metrics.

Even though we propose DialBERT ensembles through post-training method, the best model could only reach an F1 score of 45.3%, which indicates that the task of disentanglement is still a hard problem to solve.

5.2 Ablation Results

Table3 displays the ablation analysis of different components. As we discussed before, DialBERT

	VI	ARI	1-1	F1	P	R
Our model	**92.9**	68.1	**80.0**	43.9	40.5	**47.9**
- feature	92.7	**69.2**	78.4	**44.3**	**42.1**	46.7
- adaptation	92.5	67.8	78.6	41.0	37.6	45.1
- BiLSTM	90.8	62.9	75.0	32.5	29.3	36.6

Table 3: Ablation analysis of different components.

has learned enough information to differentiate messages, so linguistic features do not result in a drastic improvement. Unsurprisingly, the performance of the model drops in 5 out of 6 evaluation metrics after removal of post-training process, which indicates post-training learns useful semantic information, especially under the condition that the dataset is in a specific domain. After the removal of BiLSTM, results fall remarkably according to all evaluation metrics. In that way, the model has to make a prediction without any context consideration. As we discussed before, context is very important to disentangle a conversation. We can see from the ablation results, every modification to BERT in our model as mentioned in Table3 contributes to the final result, especially the BiLSTM component.

6 Conclusion

Conversation disentanglement is a hard problem with broad application prospects. In this paper, we propose a novel framework towards dealing with disentanglement. Different from previous work, we introduce both local and global semantics to disentangle conversations. Moreover, in order to make DialBERT perform well in the Ubuntu domain, we also post-train our model. Our model reaches state-of-the-art results on the newly published disentanglement dataset with a substantial margin in comparison to other baseline models.

References

Paige Adams and Craig Martell. 2008. Topic detection and extraction in chat. pages 581–588.

James Allan. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*, pages 1–16. Springer.

Paul M Aoki, Matthew Romaine, Margaret H Szymanski, James D Thornton, Daniel Wilson, and Allison Woodruff. 2003. The mad hatter's cocktail party: a social mobile audio space supporting multiple simultaneous conversations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 425–432. ACM.

Paul M Aoki, Margaret H Szymanski, Luke Plurkowski, James D Thornton, Allison Woodruff, and Weilie Yi. 2006. Where's the party in multi-party?: Analyzing the structure of small-group sociable talk. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 393–402. ACM.

H. Daume III and D. Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101126.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wenchao Du, Pascal Poupart, and Wei Xu. 2017. Discovering conversational dependencies between messages in dialogs. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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.

Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1179–1189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271.

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.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Tianda Li, Xiaodan Zhu, Quan Liu, Qian Chen, Zhigang Chen, and Si Wei. 2019. Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *arXiv preprint arXiv:1904.12104*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2012. Hierarchical conversation structure prediction in multi-party chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12, pages 60–69, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_ understanding_paper. pdf*.

Matthieu Riou, Soufian Salim, and Nicolas Hernandez. 2015. Using discursive information to disentangle French language chat. In *2nd Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC 2015) / Social Media at GSCL Conference 2015*, pages 23–27, Essen, Germany.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Chulaka Gunasekara Sungjin Lee Adam Atkinson Baolin Peng Hannes Schulz Jianfeng Gao Jinchao Li Mahmoud Adada Minlie Huang Luis Lastras Jonathan K. Kummerfeld Walter S. Lasecki Chiori Hori Anoop Cherian Tim K. Marks Abhinav Rastogi Xiaoxue Zang Srinivas Sunkara Raghav Gupta Seokhwan Kim, Michel Galley. 2019. The eighth dialog system technology challenge. *arXiv preprint*.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. 2019. Domain adaptive training bert for response selection.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Henghui Zhu, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Who did they respond to? conversation structure modeling using masked hierarchical transformer.

A Appendices

A.1 Related Work

Pre-trained models such as BERT(Devlin et al., 2018) have been proven to work efficiently in many NLP tasks, but there is no previous work relevant to conversation disentanglement problem that uses pre-trained model. In this work we will present our own DialBERT model, which to best of knowledge is the first attempt to adopt a pre-trained model for disentanglement.

Domain adaptation(post-training) is an effective way to improve the performance of a model on a domain-specific corpus. Domain adaptation were first researched by (Daume III and Marcu, 2006). Labeled data from both the source domain and target domain are needed for most adaptation work(Jiang and Zhai, 2007; Whang et al., 2019; Howard and Ruder, 2018; Han and Eisenstein, 2019). Recently, domain adaptation has also been used for multi-turn response selection in a retrieval-based dialog system by (Whang et al., 2019). The dataset used in our work is derived from Ubuntu Internet Relay Chat(IRC). In order to make our model capture semantics related to Ubuntu, we used the external knowledge from DSTC8 (Seokhwan Kim, 2019) to post-train our DialBERT.

Simultaneous conversation occurs not only in informal social interactions but also in conversations involving several participants in our daily lives. Aiming to separate intermingled messages into detached conversations, conversation disentanglement is the key research towards dealing with that problem. The research for conversation disentanglement dates back to (Aoki et al., 2003) which conducted a study of voice conversations among 8-10 people with an average of 1.76 activate conversations at any given time. Further research not only propose datasets(Elsner and Charniak, 2008; Mehri and Carenini, 2017; Riou et al., 2015) but also models (Mehri and Carenini, 2017; Jiang et al., 2018). In recent studies, Mehri and Carenini (2017) used recurrent neural networks(RNNs) to model adjacent messages as additional features. Jiang et al. (2018) was the first to propose to use convolutional neural networks to estimate the conversation-level similarity between closely posted messages. In recent research, Zhu et al. (2019) proposed masked hierarchical transformer based on BERT to learn a better conversation structure. Their approach need conversation structure in advance to formulate the mask for their transformer model. During the test process, they construct their mask according to previous prediction, in that way, their mask could not give reliable conversation structure information, because of the relative low accuracy of the model's prediction.

Our model could introduce both global and local conversation semantics without explicitly introduce conversation structure result in achieving state-of-the-art results by outperforming other models by wide margin gap.

	Train	dev	test
Speaker	15753	1220	1480
Conversation	17619	749	962
Messages	67463	2500	5000

Table 4: Statistic of disentanglement dataset that we used.

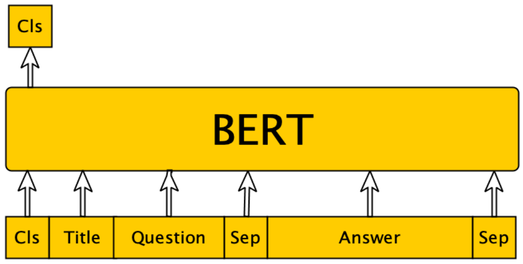

Figure 2: Adaptation process

A.2 Domain Adaptation

Ubuntu forum data is given by Google DSTC8 (Seokhwan Kim, 2019) competition track2's external source. Each sample contains a title and a question followed by an answer for that question. All questions lie within the context of Ubuntu. Different from random shuffling the external data sentences as input, we manually set the input of post-training process. As shown in figure2, we followed the input format of BERT. Specifically, we used title and question as sentence A, and answer as sentence B. For every sentence A, we randomly pick a answer as a negative sample.

A.3 Dataset

The dataset proposed by (Kummerfeld et al., 2019) is 16 times larger than all previously released datasets combined. Here is the data sample in table 5. The training set was sampled in three ways: (1) 95 uniform length samples, (2) 10 smaller samples to check annotator agreement, (3) 48 time spans of one hour, which could maintain the diversity in terms of number of messages, the number of par-

Previous message index	Index	Message
...	...	...
996	1000	[03:04] Amaranth: @cliche American
992	1001	[03:04] Xenguy: @Amaranth I thought you were – welcome mortal ;-)
1000	1002	[03:04] cliche: @ Amaranth, hahahaha
1003	1003	=== welshbyte has joined #ubuntu
997	1004	[03:04] e-sin: no i just want the normal screensavers
995	1005	[03:04] Amaranth: @benoy Do you have cygwinx installed and running?
1006	1006	[03:04] babelfishi: can anyone help me install my Netgear MA111 USB adapter?
1004	1007	[03:04] e-sin: i have a 16mb video card
1008	1008	=== regeya has joined #ubuntu
1007	1009	[03:04] e-sin: TNT2 :)
1001	1010	[03:05] Amaranth: @Xenguy hehe, i do side development
1007	1011	[03:05] jobezone: @e-sin then it's xscreensaver and xscreensave-gl for opengl ones.
1005	1012	[03:05] benoy: how do i install that? I couldn't find that in the list of things
1010	1013	[03:05] Amaranth: @Xenguy things like alacarte and easyubuntu
...	...	...

Table 5: Dataset format.

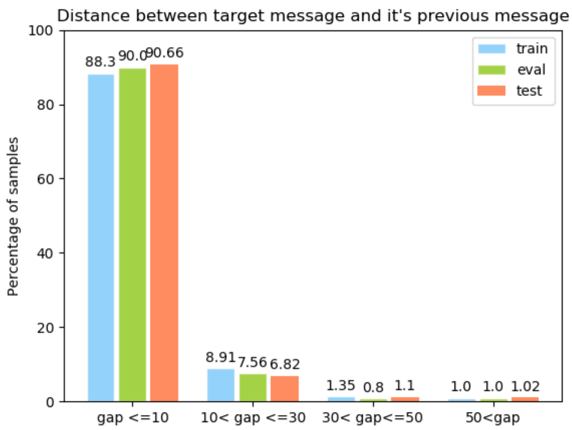

Figure 3: The percentage of Distance between target message and its preceding message

ticipants, and the percentage of directed messages. Moreover, with the use of **Adjudication**, the quality of this dataset is reliable. Practically, we will use train, dev, test folders of that dataset, excluding leave Pilot, Channel folders out. As shown in table 4, about 74963 messages were used, which is still much larger than previous work.

A.4 Training Detail

We use the base version of BERT. The initial learning rate was set to 2e-5. The maximum sequence length was set to 100. We use a batch size of 4. As shown in figure 3, around 99% target messages could find their respective preceding messages in 50 round context, so context range T was set to 50, with 49 messages before target message were considered as context. DialBERT + future context will consider 10 messages occurring after the target message. The hidden size of the Bi-LSTM component of our model was set to 384, so the concatenated output is 768 which is the same as *base-BERT*'s output. The heuristic classifier has 3072 hidden units. Dropout is implemented to the output of DialBERT and Heuristic classifier with the ratio of 0.1.

A.5 Ensemble Strategy

- **Model-AVG** In Model-AVG process, we average the weight of the model across several Dialogue BERT models.
- **Prob-AVG** In Probability-AVG process, we average the weight of the model prediction probability for each sample across different model.
- **Vote-AVG** In Vote-AVG process, we create an ensemble of DialBERT by considering the context message with most votes from each model as our prediction within the same conversation with the target message.

Specifically, different strategy need different amount of models to reach best result. For Model-AVG is 2 models. Probability-AVG is 8 models. Vote-AVG is 8 models.