# Revisiting Pre-trained Language Models for Conversation Disentanglement

**Tung-Thien Lam**
Dept. of Computer Sci. and Eng.
Yuan Ze University
Taoyuan, Taiwan
s1136058@mail.yzu.edu.tw

**Cheng-Zen Yang**
Dept. of Computer Sci. and Eng.
Yuan Ze University
Taoyuan, Taiwan
czyang@saturn.yzu.edu.tw

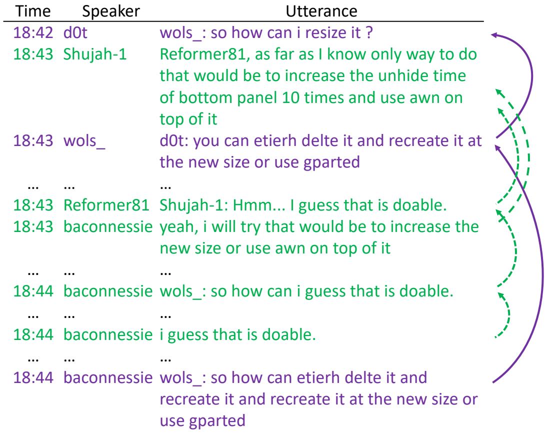
## Abstract

Multi-party conversation is a popular form in online group chatting. However, the interweaving of utterance threads complicates the understanding of the dialogues for participants. Many conversation disentanglement models have been proposed using transformer-based pre-trained language models (PrLMs). However, advanced transformer-based PrLMs have not been extensively studied. This paper investigates the effectiveness of six advanced PrLMs: BERT, XLNet, ELECTRA, RoBERTa, DeBERTa, and ModernBERT. The experimental results show that DeBERTa has outstanding performance than other PrLMs for the conversation disentanglement task.

***Keywords:*** Multi-party Conversation, Conversation Disentanglement, Pre-trained Language Models, Performance Evaluation

## 1 Introduction

Online group chatting provides important channels to multiple participants to communicate, discuss opinions, and share information. Due to its popularity, a huge amount of conversation is generated daily. Since multiple participants are present in a chatting room simultaneously, there are many different utterance threads of various topics concurrently happening in the room and they are usually intertwined without specific structural information. Although these dialogues contain abundant valuable information, the interweaving of utterance threads complicates the understanding of the dialogues for participants (Shen et al., 2006; Elsner and Charniak, 2010; Uthus and Aha, 2013). Figure 1 shows a simplified example in which only two interwoven utterance threads are illustrated. In the dialogue, the utterance threads lack coherence not only for they are intertwined but also for many irrelevant utterances appear between these threads.



| Time | Speaker | Utterance |
|---|---|---|
| 18:42 | d0t | wols_: so how can i resize it ? |
| 18:43 | Shujah-1 | Reformer81, as far as I know only way to do that would be to increase the unhide time of bottom panel 10 times and use awn on top of it |
| 18:43 | wols_ | d0t: you can etierh delte it and recreate it at the new size or use gparted |
| ... | ... | ... |
| 18:43 | Reformer81 | Shujah-1: Hmm... I guess that is doable. |
| 18:43 | baconnessie | yeah, i will try that would be to increase the new size or use awn on top of it |
| ... | ... | ... |
| 18:44 | baconnessie | wols_: so how can i guess that is doable. |
| ... | ... | ... |
| 18:44 | baconnessie | i guess that is doable. |
| ... | ... | ... |
| 18:44 | baconnessie | wols_: so how can etierh delte it and recreate it and recreate it at the new size or use gparted |

Figure 1: An example of two utterance threads extracted from the Ubuntu IRC data (Kummerfeld et al., 2019). They are expressed in purple and green.

Recently, many conversation disentanglement models (Zhu et al., 2020, 2021; Li et al., 2022; Ma et al., 2022) have proposed by employing pre-trained language models (PrLMs) such as BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2020) to improve the disentanglement performance. However, previous research has not extensively explored the effectiveness of advanced transformer-based models such as ELECTRA (Clark et al., 2020) and DeBERTa (He et al., 2021).

In this paper, six advanced transformer models are investigated for their effectiveness in conversation disentanglement: BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and ModernBERT (Warner et al., 2025). To evaluate the performance of each PrLM, we construct the disentanglement model based on the MF (Manual Features) model (Zhu et al., 2021) because the kernel of MF is relatively concise by using a 2-layer FFN (Feed-Forward Network) model. Different PrLM models can be simply employed in MF and evaluated. The experiments are conducted using the Ubuntu IRC dataset (Kummerfeld et al., 2019) because this dataset has been widely used to evaluate the disentanglement performance of different approaches.

The constructed MF-based model is a two-step disentanglement model. The first step is to perform link prediction to find the reply-to relation between a target utterance and a group of history utterances. Based on the link prediction results, the second step is to perform clustering to determine the utterance threads. The experimental results show that DeBERTa outperforms other PrLMs in terms of both link prediction metrics and clustering metrics.

The rest of the paper is organized as follows. Section 2 reviews previous studies employing PrLMs on conversation disentanglement. Section 3 describes the task definition and the dataset. Section 4 describes the studied pre-trained transformer-based models. Section 5 presents the experiments and discusses the experimental results. Finally, Section 6 concludes the paper.

## 2 Related Work

Prior studies have proposed various models for the multi-party conversation disentanglement problem (Uthus and Aha, 2013)(Uthus and Aha 2013). As pre-trained language models (PrLMs) have been widely used in natural language processing (NLP) tasks, many recent disentanglement models employ PrLMs to improve the disentanglement performance. In 2020, Zhu et al. proposed a masked hierarchical transformer model (Zhu et al., 2020) using BERT to generate feature vectors and make pairwise decisions. In 2021, Zhu et al. studied three transformer-based PrLMs with the MF model (Zhu et al., 2021): BERT (Devlin et al., 2018), ALBERT (Lan et al., 2020), and Poly-Encoder (Humeau et al., 2020). Their experimental results show that BERT combined with MF outperforms other models.

In 2022, Ma et al. proposed a BERT-based model StructBERT considering structural information of dialogues (Ma et al., 2022). In StructBERT, BERT is used to capture the contextual information of utterances. Li et al. proposed a hierarchical pre-trained model DialBERT (Li et al., 2022) using BERT to capture the matching relationship between two utterances. Jiang et al. proposed an intent-based mutual learning model MuiDial (Jiang et al., 2022) using BERT to generate utterance embeddings. In 2023, Bhukar et al. proposed an end-to-end deep reinforcement learning model (Bhukar et al., 2023) using StructBERT to get high-quality link prediction results.

In 2024, Gao et al. proposed an end-to-end implicit addressee model IAM (Gao et al., 2024) using BERT to generate utterance embeddings. Li et al. proposed a model using discourse-aware encoding and hierarchical ranking loss (DiHRL) (Li et al., 2024). As StructBERT, DiHRL uses BERT to perform contextual information encoding for utterances.

To the best of our survey, only the work of Zhu, Lau, and Qi Zhu et al. (2021) have investigated three PrLMs. This paper investigates more advanced transformer-based PrLMs that have been proposed recently.

## 3 Task Definition and Datasets

Since this work investigates the effectiveness of various PrLMs based on MF disentanglement model (Zhu et al., 2021), this paper frames the task as a problem to find reply-to relations (link prediction) and discover utterance threads (clustering). Given an utterance $u_i$ in a dialogue $D$ and a list of candidate prior utterances $\{u_j\}$ in the same dialogue, the disentanglement model firstly predicts the parent utterance of $u_i$ from $\{u_j\}$. After all reply-to relations in a segment of $D$ have been predicted, the model performs clustering to aggregate utterance threads.

To evaluate the effectiveness of PrLMs, the Ubuntu IRC dataset (Kummerfeld et al., 2019) is used because it has been widely used for performance evaluation in many studies. This dataset consists of three parts: 67,463 utterances for training, 2,500 utterances for validation, and 5000 utterances for testing.

## 4 Pre-Trained Models

To evaluate these PrLMs, we construct an MF-based model in which a pairwise model is used to predict the reply-to link relations as shown in Figure 2, where $k_h$ defines the number of utterances including the target utterance $u_i$ for reply-to relevance calculations, $w_r^t$ is the $t$-th word embedding in the $r$-th utterance, and $mf_{ij}$ represents the manually defined features including the utterance characteristics and the mutual relationships like the number of intervening messages, the word overlap ratio, and the condition of words in common. In this work, $k_h$ is set to 50. Thereafter, our MF-based

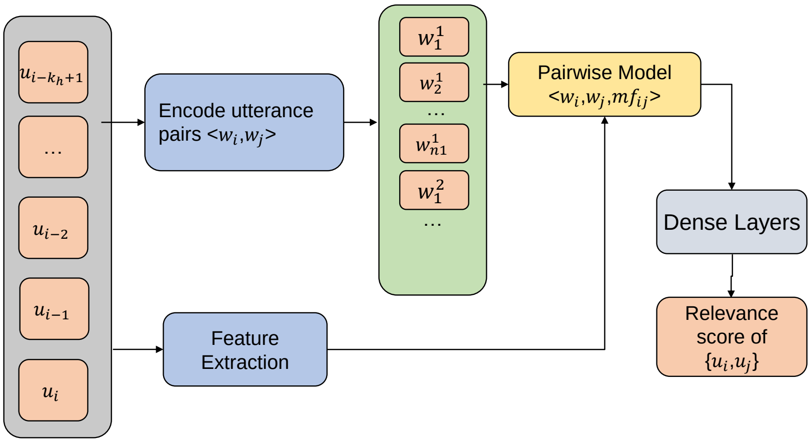

$u_{i-k_h+1}$
…
$u_{i-2}$
$u_{i-1}$
$u_i$
Encode utterance pairs $<w_i,w_j>$
$w_1^1$
$w_2^1$
…
$w_{n1}^1$
$w_1^2$
…
Pairwise Model $<w_i,w_j,mf_{ij}>$
Dense Layers
Relevance score of $\{u_i,u_j\}$
Feature Extraction

Figure 2: The pairwise model to perform link prediction in the MF-based model.

model uses the Union-Find algorithm as (Kummerfeld et al., 2019) instead of the bipartite graph algorithm used in (Zhu et al., 2021) to perform clustering because of the wide employment of Union-Find in many related studies.

In this paper, six PrLMs are investigated. They are listed as follows:

- **BERT**: This model uses Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) to read a given sentence bidirectionally. This enables BERT to capture both left and right context simultaneously, improving the understanding of semantics. It has been widely used in many disentanglement models.

- **XLNet**: This model employs an autoregressive pre-training method to learn context bidirectionally. As shown in (Yang et al., 2019), XLNet outperforms BERT on many NLP tasks.

- **ELECTRA**: This model employs a replaced token detection method instead of an MLM approach. It trains a discriminator to distinguish real tokens from the replaced ones, and uses a MLM-based generator to predict corrupted tokens. This allows ELECTRA to use all tokens of the given input for learning, gaining more computational efficiency and parameter-effectiveness than BERT.

- **RoBERTa**: This model enhances the performance by modifying several BERT design features, including removing the NSP loss and training on a larger corpus with dynamic masking.

- **DeBERTa**: This model employs two mechanisms, Disentangled Attention and Enhanced Mask Decoder, to enhance task performance. With the Disentangled Attention approach, DeBERTa represents the content and relative position information of a token into two distinct vectors. With the Enhanced Mask Decoder approach, DeBERTa considers the absolute position information of tokens in the decoding layer to capture more complementary information. Compared with RoBERTa-Large, DeBERTa can achieve better performance with less training data.

- **ModernBERT**: This model integrates modern refinements, including rotary positional embeddings, root mean square (RMS) normalization, and multi-query attention into the BERT core. Compared with BERT, ModernBERT is optimized for longer context lengths.

## 5 Experiments

We have conducted experiments to evaluate the disentanglement performance of the MF-based model using different transformer-based PrLMs. The following implementations are used for the studied PrLMs: bert-base-uncased for BERT, xlnet-base-cased for XLNet, electra-base for ELECTRA, roberta-base for RoBERTa, deberta-v3-base for DeBERTa, and ModernBERT-base for ModernBERT.

We use Precision, Recall, and F1 to measure the link prediction performance. They are defined as follows:

$$\text{Precesion} = \frac{\text{Correctly predicted links}}{\text{All predicted links}}, \quad (1)$$

$$\text{Recall} = \frac{\text{Correctly predicted links}}{\text{All true links}}, \quad (2)$$

$$\text{F1} = \frac{2 \times Precision \times Recall}{(Precision + Recall)}. \quad (3)$$

For clustering performance, we use 1-VI (Variation of Information), ARI (Adjusted Rand Index), MCP (Matched-Cluster Precision), MCR (Matched-Cluster Recall), and MCF (Matched-Cluster F1). Because VI shows the dissimilarity between two clusters, this work uses 1-VI defined as follows:

$$1 - VI = 1 - \frac{H(Y|X) + H(X|Y)}{\log(n)}, \quad (4)$$

where $X$ and $Y$ represent two utterance threads, $H()$ is the entropy function, and $n$ is the number of the utterances. ARI shows the similarity of two clusters according to the links. It is defined as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{l_i}{2} \sum_j \binom{l_j}{2}\right]}{\binom{n}{2}}}{\frac{\left[\sum_i \binom{l_i}{2} + \sum_j \binom{l_j}{2}\right]}{2} - \frac{\left[\sum_i \binom{l_i}{2} \sum_j \binom{l_j}{2}\right]}{\binom{n}{2}}}, \quad (5)$$

where $n_{ij}$ is the number of links that appear in the predicted cluster $i$ and also the true cluster $j$, $l_i$ is the number of the links in the cluster $i$, $l_j$ is the number of the links in the cluster $j$, and $n$ is the number of the ground truth links. MCP is the Precision of the exactly-matched clusters. MCR is the Recall of the exactly-matched clusters. MCF is the harmonic mean of MCP and MCR. They are defined as follows:

$$\text{MCP} = \frac{\text{Exactly matched clusters}}{\text{All predicted clusters}}, \quad (6)$$

$$\text{MCR} = \frac{\text{Exactly matched clusters}}{\text{All true clusters}}, \quad (7)$$

$$\text{MCF} = \frac{2 \times MCP \times MCR}{(MCP + MCR)}. \quad (8)$$

All models are executed 10 times with random initializations on GPUs. The results are averaged.

We use the same settings for all models without any fine-tuning. There are three hidden layers (256, 128, 64). The optimizer is AdamW. The learning rate is 5e-5. The number of epochs is 10. The loss function is CrossEntropyLoss. The dropout rate is 0.1. The batch size is 2 with a gradient accumulation of 32. The maximum number of tokens of an utterance is 60.

Table 1 shows their link prediction performance. From Table 1, we can find that DeBERTa outperforms other PrLMs for link prediction and ELECTRA takes second place. BERT continues to deliver consistent performances. However, ModernBERT does not perform well. One possible reason is that the Ubuntu IRC dialogue dataset is a kind of the QA task, and the utterance threads are interwoven. The characteristics of the Ubuntu IRC dialogue dataset hinder the performance of ModernBERT as the findings in (Antoun et al., 2025).

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BERT | 0.7277 | 0.7014 | 0.7144 |
| XLNet | 0.7209 | 0.6951 | 0.7077 |
| ELECTRA | 0.7288 | 0.7024 | 0.7153 |
| RoBERTa | 0.7281 | 0.7019 | 0.7147 |
| DeBERTa | **0.7364** | **0.7100** | **0.7230** |
| ModernBERT | 0.7219 | 0.6960 | 0.7087 |

Table 1: Link prediction performance of the MF-based model with different PrLMs.

Table 2 shows the clustering performance of each model. DeBERTa still outperforms other PrLMs. The results of Tables 1 and 2 show that DeBERTa achieves the best performance among the studied PrLMs.

| Model | 1-VI | ARI | MCP | MCR | MCF |
|---|---|---|---|---|---|
| BERT | 0.9095 | 0.6217 | 0.3323 | 0.3915 | 0.3594 |
| XLNet | 0.9064 | 0.6222 | 0.3340 | 0.3721 | 0.3518 |
| ELECTRA | 0.9123 | 0.6543 | 0.3382 | 0.3865 | 0.3606 |
| RoBERTa | 0.9132 | 0.6424 | 0.3312 | 0.3986 | 0.3616 |
| DeBERTa | **0.9175** | **0.6644** | **0.3656** | **0.4175** | **0.3897** |
| ModernBERT | 0.9068 | 0.6128 | 0.3273 | 0.3837 | 0.3529 |

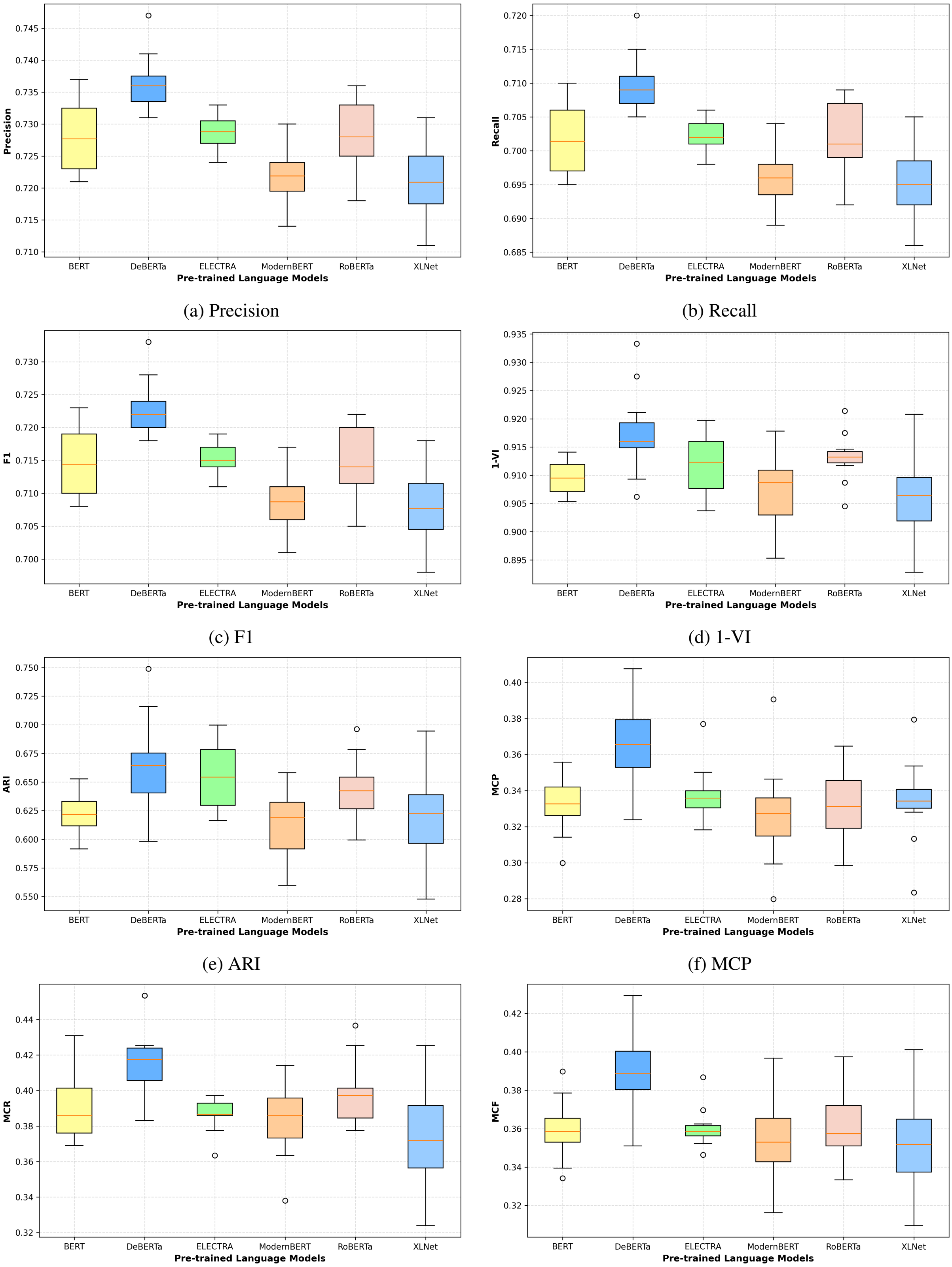
Table 2: Clustering performance of the MF-based model with different PrLMs.

Figure 3 shows the boxplot of the performance of the studied PrLMs in terms of the investigated metrics in the experiments. As shown in Figure 3, DeBERTa also has the best median scores on all performance metrics among the studied PrLMs.

## 6 Conclusions

Multi-party conversation is a popular form to discuss opinions, share information, and discover solutions for problems. However, the interweaving of utterance threads complicates the understanding of the dialogues for participants.

In the past, many conversation disentanglement models have been proposed using transformer-based PrLMs. However, advanced transformer-based PrLMs have not been extensively investigated.



(a) Precision

(b) Recall

(c) F1

(d) 1-VI

(e) ARI

(f) MCP

(g) MCR

(h) MCF

Figure 3: Boxplot of performance of the MF-based model with different PrLMs.

In this paper, six advanced transformer-based models are investigated for their effectiveness in conversation disentanglement: BERT, XLNet, ELECTRA, RoBERTa, DeBERTa, and ModernBERT. The experimental results show that DeBERTa outperforms other PrLMs for the conversation disentanglement task.

There are still some issues to be investigated further in the future. Firstly, our study does not discuss the best performance of each model because we use the same settings for all models without any finetuning. Therefore, more extensive investigations will be conducted to explore the best performance of these models. Secondly, the number of the studied PrLMs is limited. In the future, other advanced PrLMs will be included in the investigation. Finally, the MF-based model considers the manually defined features that are extracted for the Ubuntu IRC dataset. Other disentanglement models with better generalizability will be considered for more comprehensive analysis on PrLMs.

## Acknowledgments

This work is supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC 113-2221-E-155-048-MY2. The authors would also like to express their sincere thanks to anonymous reviewers for their precious comments.

## References

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2025. ModernBERT or DeBERTaV3? Examining Architecture and Data Influence on Transformer Encoder Models Performance. *CoRR*, arXiv:2504.08716.

Karan Bhukar, Harshit Kumar, Dinesh Raghu, and Ajay Gupta. 2023. End-to-End Deep Reinforcement Learning for Conversation Disentanglement. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI-23)*, volume 37, pages 12571–12579.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. 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 (NAACL 2018)*, pages 4171–4186.

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

Jingsheng Gao, Zeyu Li, Suncheng Xiang, Zhuowei Wang, Ting Liu, and Yuzhuo Fu. 2024. Toward an End-to-End Implicit Addressee Modeling for Dialogue Disentanglement. *Multimedia Tools and Applications*, 83(28):70883–70906.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021)*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Ziyou Jiang, Lin Shi, Celia Chen, Fangwen Mu, Yumin Zhang, and Qing Wang. 2022. MuiDial: Improving Dialogue Disentanglement with Intent-Based Mutual Learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 4164–4170.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-seng Chua, and Donghong Ji. 2024. Revisiting Conversation Discourse for Dialogue Disentanglement. *ACM Transactions on Information Systems*, 43(1).

Tianda Li, Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2022. Conversation- and Tree-Structure Losses for Dialogue Disentanglement. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 54–64.

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.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2022. Structural Characterization for Dialogue Disentanglement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–297.

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 (SIGIR 2006)*, pages 35–42.

David C. Uthus and David W. Aha. 2013. Multiparticipant Chat Analysis: A Survey. *Artificial Intelligence*, 199–200:106–121.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Henghui Zhu, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Who Did They Respond to? Conversation Structure Modeling Using Masked Hierarchical Transformer. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, 05, pages 9741–9748.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2021. Findings on Conversation Disentanglement. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association (ALTA 2021)*, pages 1–11.