Learning-by-Narrating: Narrative Pre-Training for Zero-Shot Dialogue Comprehension

Chao Zhao[1*] **Wenlin Yao**[2] **Dian Yu**[2]
Kaiqiang Song[2] **Dong Yu**[2] **Jianshu Chen**[2]
zhaochao@cs.unc.edu
{wenlinyao,yudian,riversong,dyu,jianshuchen}@tencent.com
[1] UNC Chapel Hill, Chapel Hill, NC [2] Tencent AI Lab, Bellevue, WA

Abstract

Comprehending a dialogue requires a model to capture diverse kinds of key information in the utterances, which are either scattered around or implicitly implied in different turns of conversations. Therefore, dialogue comprehension requires diverse capabilities such as paraphrasing, summarizing, and commonsense reasoning. Towards the objective of pre-training a zero-shot dialogue comprehension model, we develop a novel narrative-guided pre-training strategy that *learns by narrating* the key information from a dialogue input. However, the dialogue-narrative parallel corpus for such a pre-training strategy is currently unavailable. For this reason, we first construct a dialogue-narrative parallel corpus by automatically aligning movie subtitles and their synopses. We then pre-train a BART model on the data and evaluate its performance on four dialogue-based tasks that require comprehension. Experimental results show that our model not only achieves superior zero-shot performance but also exhibits stronger fine-grained dialogue comprehension capabilities. The data and code are available at https://github.com/zhaochaocs/Diana.

1 Introduction

Dialogue comprehension (Sun et al., 2019; Cui et al., 2020) aims to capture diverse kinds of key information in utterances, which are either scattered around or implicitly implied in different turns of conversations. Therefore, it requires different capabilities such as paraphrasing (Falke et al., 2020), summarizing (Gliwa et al., 2019), and commonsense reasoning (Arabshahi et al., 2021). Recent advances in pre-trained language models (PLMs) (Devlin et al., 2019; Radford et al., 2019) have been applied to the problem (Jin et al., 2020; Liu et al., 2021). However, these PLMs are generally pre-trained on formal-written texts, which are different from dialogue data in nature. Specifically, dialogues are composed of colloquial languages from multi-speakers, and utterances usually have complex discourse structures (Afantenos et al., 2015). Therefore, applying these models directly to dialogue comprehension, especially in low-resource settings, is sub-optimal.

To learn better dialogue representations, recent studies have designed several dialogue-specific pre-training objectives such as speaker prediction (Qiu et al., 2021), utterance prediction (Chapuis et al., 2020), response selection (Wu et al., 2020), and turn order restoration (Zhang and Zhao, 2021). These methods, albeit improve over the vanilla PLMs, usually rely on surface-level dialogue information. In particular, they still fail to train the models to explicitly learn the aforementioned capabilities which are critical for dialogue comprehension (e.g., linguistic knowledge, world knowledge, and commonsense knowledge). Furthermore, it was not able to incorporate knowledge beyond dialogue (e.g., non-verbal communications between speakers, as well as time and location information), which are also crucial for dialogue comprehension.

To pre-train a zero-shot dialogue comprehension model with the aforementioned features, we develop a novel generative pre-training strategy that *learns by narrating* the key information from a dialogue input (see Figure 1 for an example). In particular, the generated narrative text is supposed to not only (i) paraphrase the gists of the dialogue but also (ii) carry certain inferred information (e.g., the time and location of a scene and relations between speakers) that are not explicitly mentioned in the dialogues. Learning to narrate such information helps the model to learn varied lexical, syntactic, and semantic knowledge of dialogue. It also enhances the model's ability to infer extra information beyond the literal meaning within dialogues, which will benefit the model's capability of dialogue comprehension.

[*] Work was done during the internship at Tencent AI lab.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 212 - 218

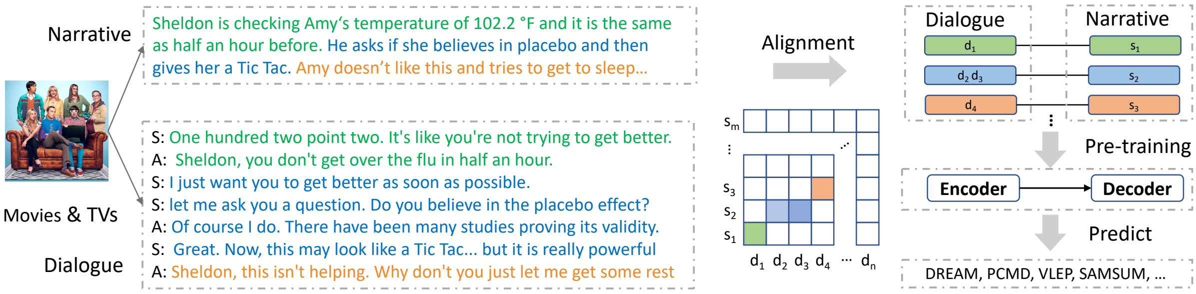

Figure 1: Overview of the *learning-by-narrating* strategy for pre-training a zero-shot dialogue comprehension model (with an encoder-decoder architecture).

However, the *learning-by-narrating* strategy would require a dialogue-narrative parallel corpus, which, to our best knowledge, is not publicly available. For this reason, we first create **DIANA**, a large-scale dataset with (**DIA**logue, **NA**rrative) pairs automatically collected from subtitles of movies and their corresponding plot synopses. We consider dialogues from movie subtitles as they are close to daily human-to-human conversations (Zhang and Zhou, 2019). In addition, the movie synopses include rich narrative information, which is helpful for dialogue comprehension. After data collection and strict quality control, we obtain a dataset with 243K (dialogue, narrative) pairs written in English. As the automatic data construction procedure is language-independent, it can be applied to low-resource languages as well.

We then pre-train a BART model (Lewis et al., 2020) on the constructed corpus with the proposed *learning-by-narrating* strategy, and evaluate it on four dialogue-based tasks that require comprehension. In zero-shot settings, our pre-trained model outperforms the BART baseline on all tasks by a large margin (e.g., $+8.3\%$ on DREAM (Sun et al., 2019)), demonstrating the success of our approach.

The contributions of this paper are three-fold:

- We propose a novel *learning-by-narrating* pre-training strategy for dialogue comprehension;
- We release **DIANA**, a new large-scale dialogue-narrative parallel corpus;
- Experiments show that our pre-trained dialogue comprehension model achieves superior zero-shot performance on a variety of downstream tasks.

2 DIANA: A Dialogue-Narrative Corpus

In this section, we describe the procedure to create the dialogue-narrative parallel dataset.

2.1 Data Collection and Segmentation

We collect 47,050 English subtitles of movies and TV episodes released from Opensubtitle (Lison et al., 2018) and their corresponding synopses from online resources such as Wikipedia and TMDB. To link the subtitle and synopsis of the same movie or TV episode, we require a subtitle and a synopsis to have the same title and the release year, as well as a high overlap rate ($> 50\%$) on role names.

The subtitle and synopsis of a movie are too long for a PLM. To facilitate pre-training, we split both the subtitle and synopsis into smaller segments and align the related segments from each part to shorter (dialogue, narrative) pairs. We split subtitles using the time interval δ_T between utterances and split a synopsis into sentences. We set $\delta_T = 5s$.

2.2 Data Alignment

We aim to align the dialogue sessions $\{d_1, \ldots, d_n\}$ and narrative segments $\{s_1, \ldots, s_m\}$ with maximum global similarity to form (dialogue, narrative) pairs. For each dialogue session d_j, the goal is to find its corresponding narrative segment s_i.

Inspired by (Tapaswi et al., 2015) in which the narrative in a synopsis follows the timeline of a movie or a TV episode, we develop a dynamic time warping method to find the globally optimal alignment score. During aligning, some narrative segments contain information beyond the dialogue, so they cannot be aligned to any dialogue session. We therefore allow our algorithm to skip at most k narrative segments during alignment searching:

$$\mathcal{A}(i,j) = \max_{0 \le k \le K+1} \mathcal{A}(i-k, j-1) + \mathcal{S}\left(s_i, d_j\right), \quad (1)$$

where $\mathcal{A}(i,j)$ denotes the optimal alignment score of the first i narrative segments and the first j dialogue sessions. $\mathcal{S}\left(s_i, d_j\right)$ is the text similarity between s_i and d_j.

We compare the performance of three text similarity measures: Jaccard similarity, Rouge-1F, and

Similarity Function	Accuracy
Jaccard	57.98
Rouge-1F	60.01
TF-IDF	67.20
TF-IDF normalized	**71.95**

Table 1: Alignment accuracy of different similarity measures on MovieNet.

TF-IDF. In consideration of time efficiency, we don't apply more advanced neural methods. We compare these similarity measures on MovieNet dataset (Huang et al., 2020), which provides a manual alignment between the segments of subtitles and synopses of 371 movies. [1] We evaluate the performance of each similarity measure by alignment accuracy, a.k.a, the percentage of dialogue sessions that are correctly aligned to the corresponding narrative segment. As shown in Table 1, TF-IDF performs best among all similarity measures. We also find that a narrative-wise L_2 normalization of the TF-IDF can further improve the alignment accuracy. It helps to penalize the similarity of (d_j, s_i) when s_i has high similarity with many dialogues (e.g., when s_i contains common words or protagonists' names.) We therefore choose the normalized TF-IDF as our similarity function. We further analyze the errors during alignment and find that 85.94% of errors happen because the dialogue session is aligned to the previous or next segment of the gold narrative segment. It indicates that most of the errors happen locally. Figure 2 shows an example from MovieNet, where the red line and the blue line indicate the gold alignment and the predicted alignment via normalized TF-IDF, respectively. It shows that the two lines are generally well overlapped except for some local discrepancies.

2.3 Quality Control

After data alignment, each narrative segment s_i can be aligned to multiple dialogues. To consider the local alignment errors, we also merge the aligned dialogues of s_{i-1} and s_{i+1} to the dialogues of s_i. Some of these dialogues may not be relevant to s_i. To select the relevant dialogues, we use a greedy method to incrementally select dialogues until the rouge-F score between the narrative and the selected dialogues doesn't increase. After selection, we concatenate the selected dialogues and preserve their relative position. We finally obtain around 1.5 Million (dialogue, narrative) pairs.

[1] We use MovieNet for test purposes only.

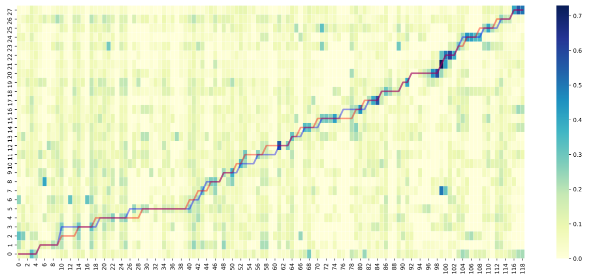

Figure 2: The Alignment of dialogues and narrative segments of a movie. X-axis and Y-axis are the ID of dialogue sessions and narrative segments, respectively. The variety of colors depicts the different similarity values between a dialogue session and a narrative segment. The blue line is the predicted alignment via normalized TF-IDF while the red line is the gold alignment.

To further improve the quality of data, we filter out pairs where the dialogue and the narrative are irrelevant. To evaluate the relevance, we use two automatic measures: Coverage and Density (Grusky et al., 2018). Low Coverage and Density indicate that the narrative text is either too abstractive or irrelevant to the dialogue. We thus only select the pairs with Coverage > 0.5 and Density > 1. After this strict quality control, we obtain 243K (dialogue, narrative) pairs as the final DIANA dataset, which is a high-quality subset of the original dataset. The average length of the dialogue and the narrative are 58 tokens and 18 tokens, respectively.

2.4 Analysis of Knowledge Type

To analyze what types of knowledge are included in DIANA, we randomly sample 100 instances and manually categorize the relation between dialogue and the corresponding narrative text into seven knowledge types. We show the percentage of each knowledge type in parentheses and in Figure 3 as well. The knowledge types are:

- **Summarizing** (39%): The narrative text summarizes multiple utterances as a concise statement to reflect the salient event or information of the dialogue.
- **Visual/Audial** (17%): The narrative text provides extra visual or audial information of the dialogue, such as the location of the dialogue, the speakers' actions, and ambient sounds.
- **Paraphrasing** (14%): The narrative text restates speakers' utterances using other words.
- **Text Matching** (9%): The narrative text is directly copied from the utterances of speakers.

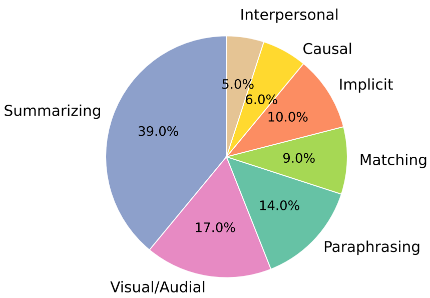

Figure 3: The knowledge type distribution in DIANA.

- **Implicit** (10%): The narrative text provides extra information that is not explicitly mentioned in the dialogue.
- **Causal** (6%): The narrative text describes the cause and effect relationship between events.
- **Interpersonal** (5%): The narrative text reveals the relationships between speakers.

Among these knowledge types, *Summarizing* and *Visual/Audial* are the two most frequent ones. They are followed by *Paraphrasing* and *Text Matching*, which contribute to 23% in total. It also shows that narratives use paraphrasing more often than copying. Additionally, DIANA contains three higher-level knowledge types that require the awareness of real-world commonsense and more complicated inference such as implicit knowledge, causal relationships, and interpersonal relationships. The diverse knowledge types in DIANA indicate the benefit of this dataset for dialogue comprehension and other downstream tasks as well.

3 Pre-training: Learning-by-Narrating

During pre-training, we aim to inject the knowledge contained in DIANA into pre-trained models. One option is to ask the model to distinguish between a correct narrative and an incorrect narrative via a classification objective. However, it requires carefully designing additional non-trivial negative (dialogue, narrative) pairs. Therefore, we propose to directly generate a narrative text from the given dialogue by maximizing the generative probability:

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = \prod_{t=1}^{|\mathbf{y}|} p\left(y_t \mid y_{1:t-1}, \mathbf{x}; \boldsymbol{\theta}\right), \quad (2)$$

where $\mathbf{x}$ are dialogue texts and $\mathbf{y}$ are narrative texts.

There are two main advantages of using the generative objective. First, it can fully leverage the narrative information from each token of the narrative text with no need to construct negative pairs. Second, the pre-trained model can be directly applied to both generative and discriminative downstream tasks without further fine-tuning. For discriminative tasks, we calculate the probability of each candidate according to Equation 2 and choose the most probable candidate as the predicted answer.

4 Experiments

In this section, we evaluate the performance of the pre-trained model on four downstream tasks that require dialogue comprehension.

4.1 Setting

We use BART, a state-of-the-art sequence-to-sequence model, as our baseline model.[2] We use its released checkpoint and further pre-train the model on DIANA. During pre-training, we concatenate the utterances as the input and update the parameters to maximize the probability of the corresponding narrative. We use Adam as the optimizer, and we set the learning rate and weight decay to 3×10^{-5} and 0.01, respectively. Following previous studies that suggest that a larger batch size helps pre-training, we set the batch size to 1024 and pre-train the model for 1,000 steps.

4.2 Tasks

We evaluate our model's ability of dialogue comprehension on four downstream tasks. **DREAM** (Sun et al., 2019) aims to read a dialogue and select the correct answer from options of a dialogue-related question. To make the task similar to our pre-training task, we follow previous work (Chen et al., 2021) to train a T5 model to convert each (question, answer) pair to a statement. **PCMD** (Ma et al., 2018) is a passage completion task. Given a dialogue and a passage that describes the dialogue, a query is created by replacing a character mention with a variable x, and the model needs to recover the character mention. **VLEP** (Lei et al., 2020) aims to select the most probable future event given the dialogue of the current event and two candidates of future events. **SAMSum** (Gliwa et al., 2019) is a dialogue summarization task to create a concise abstractive summary for a dialogue. The first three are discriminative tasks, and SAMSum is a generative task. None of the source dialogues in these tasks are included in DIANA.

[2]We also tried T5 and Pegasus in our early experiments but did not observe better performance compared with BART.

	Data	Task	DREAM ACC	PCMD ACC	VLEP ACC	SAMSum R1	R2	RL
BART-FT	-	-	62.56	75.89	65.07	49.18	24.47	47.12
GPT-2	-	-	41.99	45.02	54.58	10.83	0.74	11.68
RoBERTa	-	-	45.22	46.25	52.28	-	-	-
BART	-	-	45.07	46.07	54.26	29.92	9.58	28.54
	DIAL	DE	46.69	47.34	55.98	30.08	9.52	29.36
	CNN	CLS	50.46	49.27	55.53	-	-	-
	CNN	GEN	52.72	45.34	58.13	31.33	9.08	28.03
	CRD3	GEN	52.96	45.71	57.12	27.07	9.09	27.64
Narrator	DIANA	GEN	**53.41**	**54.88**	**58.90**	**37.27**	**13.23**	**36.12**

Table 2: Results on four dialogue-based tasks. For models that require further pre-training, we list the corresponding pre-training dataset and task.

We evaluate the model performance on these tasks under the zero-shot setting. For discriminative tasks, we convert each test instance with K answer candidates as K (dialogue, narrative) pairs. Given the dialogue as input, we evaluate the conditional probability of each narrative according to Equation 2 and choose the most probable narrative as the predicted answer. We use accuracy (ACC) as the evaluation metric for discriminative tasks and ROUGE for the summarization task.

We compare our pre-trained model (Narrator) with strong pre-trained baselines such as GPT-2, RoBERTa, and BART. To investigate the impact of the pre-training objective, we compare with 1) BART-DIAL-DE: the original BART de-noising objectives, which is trained on the dialogue part of DIANA; and 2) BART-CNN-CLS: a classification objective, which is trained using the CNNDM dataset (See et al., 2017) to distinguish between positive and negative summaries based on the documents. Negative summaries are obtained from DocNLI (Yin et al., 2021) by replacing the words, entities, and sentences of positive summaries. We also investigate the quality of DIANA by comparing it with two large summarization datasets: CNNDM and CRD3 (Rameshkumar and Bailey, 2020). We pre-train BART to generate the summaries of these datasets from the corresponding documents and refer to the models as BART-CNN-GEN and BART-CRD3-GEN. Besides the zero-shot models, we list the supervised results finetuned on BART (BART-FT) as a reference for the upper bound.

4.3 Results

Results are shown in Table 2. Our observations are as follows. (i) When compared with vanilla PLMs, Narrator outperforms GPT-2, RoBERTa, and BART, demonstrating that the learning-by-narrating pre-training objective can improve the model's ability of dialogue comprehension. (ii) When compared with different pre-training tasks, Narrator outperforms BART-DIAL-DE, and BART-CNN-GEN outperforms BART-CNN-CLS. This indicates that the narrative-guided generative objective is more effective than the de-noising objective and the discriminative objective. (iii) When compared with different pre-training data, Narrator achieves better performance on all tasks compared with BART-CNN-GEN and BART-CRD3-GEN, demonstrating that DIANA is a more helpful resource for dialogue comprehension.

Question Type	BART	Narrator
Paraphrase+Matching	58.4	66.1 (+7.7)
Reasoning	42.2	46.2 (+4.0)
Summary	51.1	53.4 (+2.3)
Logic	43.8	48.2 (+4.4)
Commonsense	37.8	41.9 (+4.1)
Arithmetic	23.8	23.8 (+0.0)

Table 3: Accuracy by question types on DREAM.

We further analyze what types of knowledge are enhanced during pre-training. To this end, we test Narrator on a subset of the DREAM test set, which includes annotated knowledge types released along with the DREAM dataset. As shown in Table 3, compared with the vanilla BART, Narrator achieves better performance on all knowledge types except Arithmetic, which is not covered in DIANA. The performance gain indicates that the narrative pre-training contributes the most to the knowledge related to paraphrasing and matching. It also benefits from other knowledge types that require various reasoning abilities such as commonsense reasoning and logic reasoning.

5 Conclusion

We propose a *learning-by-narrating* strategy to pre-train a zero-shot dialogue comprehension model. We first construct a dialogue-narrative dataset named DIANA, which contains 243K (dialogue, narrative) pairs obtained by automatically aligning movie subtitles with their corresponding synopses. We then pre-train a dialogue comprehension model based on DIANA and evaluate its performance on four downstream tasks that require dialogue comprehension abilities. Experiments show that our model outperforms strong pre-trained baselines, demonstrating that the learning-by-narrating strategy is a promising direction for dialogue comprehension. We also hope that DIANA will promote future research in related areas.

References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. 2021. Conversational neuro-symbolic commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4902–4911.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

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

Tobias Falke, Markus Boese, Daniil Sorokin, Caglar Tirkaz, and Patrick Lehnen. 2020. Leveraging user paraphrasing behavior in dialog systems to automatically collect annotations for long-tail utterances. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 21–32, Online. International Committee on Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-Tür. 2020. MMM: multi-stage multi-task learning for multi-choice reading comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8010–8017. AAAI Press.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. 2021. A graph reasoning network for multi-turn response selection via customized pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13433–13442.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. 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 2039–2048, New Orleans, Louisiana. Association for Computational Linguistics.

Yao Qiu, Jinchao Zhang, and Jie Zhou. 2021. Different strokes for different folks: Investigating appropriate further pre-training approaches for diverse dialogue tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2318–2327, Online and Punta Cana, Dominican Republic.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. 2015. Aligning plot synopses to videos for story-based retrieval. *International Journal of Multimedia Information Retrieval*, 4(1):3–16.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Leilan Zhang and Qiang Zhou. 2019. Automatically annotate tv series subtitles for dialogue corpus construction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1029–1035. IEEE.

Zhuosheng Zhang and Hai Zhao. 2021. Structural pre-training for dialogue comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5134–5145, Online. Association for Computational Linguistics.