Combining Argumentation Structure and Language Model for Generating Natural Argumentative Dialogue

Koh Mitsuda Ryuichiro Higashinaka Kuniko Saito
NTT Human Informatics Laboratories, NTT Corporation, Japan
{koh.mitsuda.td,ryuichiro.higashinaka.tp,kuniko.saito.ku}@hco.ntt.co.jp

Abstract

Argumentative dialogue is an important process where speakers discuss a specific theme for consensus building or decision making. In previous studies for generating consistent argumentative dialogue, retrieval-based methods with hand-crafted argumentation structures have been used. In this study, we propose a method to generate natural argumentative dialogues by combining an argumentation structure and language model. We trained the language model to rewrite a proposition of an argumentation structure on the basis of its information, such as keywords and stance, into the next utterance while considering its context, and we used the model to rewrite propositions in the argumentation structure. We manually evaluated the generated dialogues and found that the proposed method significantly improved the naturalness of dialogues without losing consistency of argumentation.

1 Introduction

Argumentative dialogue is an important process where speakers discuss a specific theme for building consensus or making decisions (Toulmin, 1958; Walton, 2013). The method to automatically generate argumentative dialogues not only contributes to the realization of such a dialogue system but can also provide us with content that can give us insights regarding the theme.

In previous studies in argumentation generation, retrieval-based methods with a hand-crafted argumentation structure consisting of propositions written in natural sentences were used for generating consistent argumentative dialogue (Sato et al., 2015; Rakshit et al., 2017; Higashinaka et al., 2017; Rach et al., 2018; Sakai et al., 2020). However, these methods output propositions as utterances as they are; thus the previous context is not considered, making the generated dialogue less coherent. In addition, although generation-based methods for argumentation with language models have also been proposed, generation of natural and consistent dialogue has never been investigated (Hua and Wang, 2018; Park et al., 2019; Hidey and McKeown, 2019; Mitsuda et al., 2019).

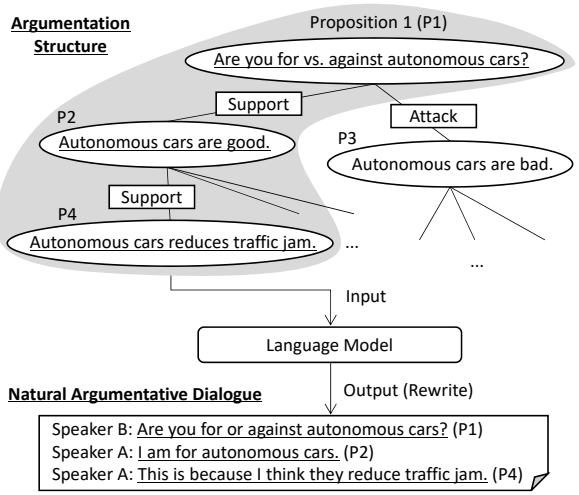

Figure 1: Our goal is to generate natural argumentative dialogue from an argumentation structure

In this study, we propose a method to generate natural argumentative dialogue by combining an argumentation structure and a language model as illustrated in Figure 1. Specifically, we propose a method to rewrite propositions of an argumentation structure into natural utterances. The method generates natural utterances on the basis of the context of dialogue and propositions' key information, such as keywords and a stance. We manually evaluated the generated dialogues and found that the proposed method significantly improved the naturalness of dialogues without losing consistency of argumentation.

2 Related Work

Our approach is related to retrieval-based generation, which generates responses by referring to the examples retrieved from resources, and keyword-

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 65–71
November 20–23, 2022.

based generation, which ensures that specified contents such as keywords are included in generated answers.

Retrieval-based generation has been applied to a wide range of tasks, such as question answering (Lee et al., 2019; Izacard and Grave, 2020), dialogue modeling (Weston et al., 2018; Roller et al., 2020), and story generation (Xu et al., 2020) in addition to argumentation generation. Our work is different from these studies in that we aim to improve the naturalness of argumentative dialogue while maintaining consistency by using a language model with argumentation structures. As far as we know, no previous work has tackled the problem of generating argumentative dialogue by using both pre-trained language models and argumentation structures.

Keyword-based generation is proposed for introducing contents specified with keywords into generated utterances (Mou et al., 2016). In addition to the content's keywords, the methods have been proposed for controlling an utterance topic by incorporating an emotional keyword (Zhou et al., 2018) and topical keywords (Xing et al., 2017). In addition, the methods have also been proposed for generating an utterance that exactly includes the given keywords (Zhu et al., 2019; Xu et al., 2020). These studies do not focus on argumentation in which logical consistency and the stance of a speaker need to be considered.

3 Datasets

We first briefly present the datasets we use in this study. We use two kinds of datasets: argumentation structure (Sakai et al., 2018) and argumentative dialogue corpus (Higashinaka et al., 2017). The argumentation structure is the source of the argumentative dialogue, which is a tree-like structure of logically connected propositions. The argumentative dialogues are used for fine-tuning the language model to rewrite the propositions into utterances while considering their previous context. The datasets are in Japanese.

The argumentation structure contains propositions in a specific theme (e.g., "Are you for vs. against autonomous cars?") as shown at the top of Figure 1 (Sakai et al., 2018). The resource is a tree where each proposition corresponds to a node written in a natural sentence and its relationships correspond to edges. The argumentation structure is constructed in five argumentation themes. The depth of trees is six, and each tree has 2,255 nodes on average.

The argumentative dialogue corpus was constructed by Higashinaka et al. (2017) in the same five themes as the argumentation structure. Speakers took opposite stances (e.g., for or against) and conducted argumentation to persuade their counterpart. They did not refer to the argumentation structure; thus there is no exact correspondence between an utterance in the dialogue and a proposition in the argumentation structure. Since each speaker has a stance, each utterance of a speaker is regarded as having the stance of that speaker. In addition, Higashinaka et al. (2017) manually labeled the argumentation-related dialogue acts (assertion, question, concession, retraction, and other) to each utterance in the corpus. The corpus has 250 dialogues (17,804 utterances in total and 71 utterances per dialogue).

4 Proposed Method

Our idea for generating an argumentative dialogue is to first create a scenario on the basis of the graph (a sequence of propositions) and then convert that graph into an argumentative dialogue. The problem is how to convert each proposition into a naturalistic utterance. For this, we use keywords-based generation in which we utilize key information about a proposition to generate an utterance. Through the investigation of the datasets, we identified the following key information.

(1) Stance of the proposition
(2) Dialogue act
(3) Turn number to indicate the depth of argumentation
(4) Keywords in the proposition

Figure 2 shows the proposed method to generate natural argumentative dialogue by combining the argumentation structure and language model. We first fine-tuned a pre-trained encoder-decoder language model with the argumentative dialogue corpus so that it can rewrite a proposition of an argumentation structure into the next utterance on the basis of its key information (stance, dialogue act, turn, and keywords) while considering context. Then, we utilized the fine-tuned language model to rewrite propositions in the argumentation structure for generating the argumentative dialogue.

In fine-tuning the model, the context before each utterance is used as input, and each utterance

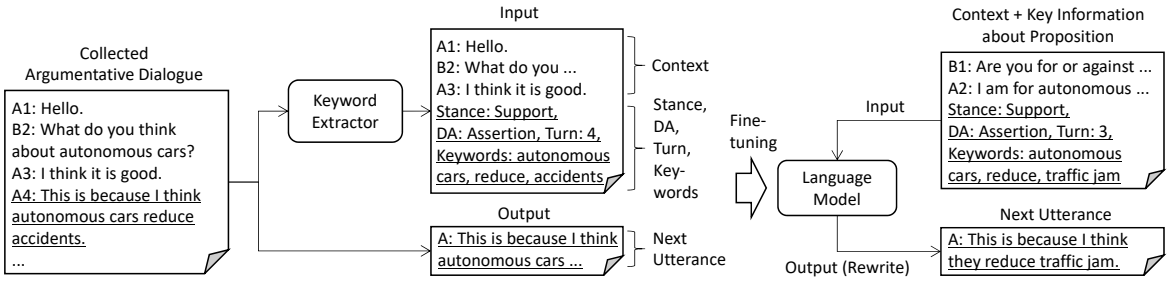

Figure 2: Proposed method to generate natural argumentative dialogue by combining argumentation structure and language model. The left part shows the process of fine-tuning a language model, and the right part shows the generation of dialogue from the argumentation structure. The language model is fine-tuned so that it can rewrite key information, such as keywords and stance, into the next utterance while considering the context. The model is applied to rewrite propositions in the argumentation structure. Input and output in the right part are the same as the examples in Figure 1.

is used as output. The model is fine-tuned so that, given the context and key information of the output utterance, the model can reconstruct the utterance. This is in the hope that when the same information is given from a proposition, a natural utterance for the proposition can be generated. A full example of an input used in Figure 2 is the following.

```
Autonomous Cars:[SEP][SPK1]Hello.[SEP]
[SPK2]What do you think about autonomous
cars?[SEP][SPK1]I think it is good.[SEP]
Stance:Support,DA:Assertion,Turn:04[SEP]
Keywords:autonomous,cars,reduce,accidents
```

Each element is divided with a separator [SEP]. The first element shows an argumentation theme. The context including three utterances at maximum follows. Then, a stance, dialogue act, turn number, and keywords are listed. We used the stance, dialogue act, and turn number labeled in the argumentative dialogue corpus for creating the training data in fine-tuning the model. The keywords are automatically extracted through a keyword extractor where a part-of-speech tagger is applied to an utterance in a dialogue in order to obtain all content words as keywords. The insertion of these kinds of information seems simple but has been reported to be effective in previous studies (Niu and Bansal, 2018; Raffel et al., 2020; Reynolds and McDonell, 2021).

The argumentative dialogue is generated in the following manner. First, by randomly selecting the path of an argumentation structure, we create a sequence of propositions as a source scenario of generated argumentative dialogue (Sakai et al., 2020). Then, the model rewrites the proposition into an output utterance using the fine-tuned language model from the top proposition to the bottom one. The generated utterance is added to the context for generating the next utterance from the next proposition. Note that the keywords are extracted from the proposition with the same keyword extractor used in fine-tuning. The dialogue act, stance, and turn number are predetermined by the scenario; each speaker's stance is fixed (e.g., a speaker A for autonomous cars and a speaker B against it) and dialogue act is determined by heuristic rules to realize a typical flow of argumentation (e.g., the first utterance is question and the second one is assertion) as will be explained in Section 5.2.

5 Experiments

We manually evaluated the dialogues generated from the proposed method. We conducted a static evaluation of dialogues by crowdsourcing, which is often used to evaluate dialogue generation in dialogue systems (Li et al., 2019).

5.1 Comparison Methods

We prepared four methods including not only the proposed method described in Section 4 (**Proposed**) but also three comparison methods (**Vanilla**, **Ret-Rewrite**, and **Kwd-Rewrite**).

(a) Vanilla: This method outputs the input sequence of propositions as it is without rewriting it by a language model. Note that, to improve the naturalness of each proposition, a Japanese sentence-end converter (Miyazaki et al., 2015) is used to normalize a phrase at the end of the proposition.

(b) Ret-Rewrite: This is a retrieval-based rewriting method that generates the next utterance from a given context and proposition. To this end, for fine-tuning the model, it is necessary to prepare

input-output pairs <context + proposition, next utterance> from the argumentative dialogue corpus and argumentation structures. Therefore, we prepared such pairs by retrieving the proposition most similar to each next utterance from the corresponding argumentation structure. For retrieving the proposition, Sentence-BERT (Reimers and Gurevych, 2019)[1] is used to calculate the similarity between a candidate proposition and the next utterance.

(c) Kwd-Rewrite: This is a keyword-based generation method without using other key information (stance, turn number, and dialogue act). This method is prepared to investigate the effectiveness of using only the extracted keywords.

For the base encoder-decoder language model, we used the Japanese version of BlenderBot (Roller et al., 2020) trained by Sugiyama et al. (2021) (the number of parameters is 1.6B).

5.2 Experimental Procedure

With regards to the evaluation protocol, we first automatically created scenarios from the argumentation structures. Then, the created scenarios were rewritten into dialogues by the proposed method for evaluation. The original scenarios are created in a manner similar to the method of Sakai et al. (2020). We conceived the following requirements for generating scenarios, which we think follow a general argumentation flow.

(1) Speakers A and B first assert their stance (e.g., for or against autonomous cars).
(2) One speaker (e.g., A) supports his/her stance with a proposition.
(3) The other speaker (e.g., B) counters with a proposition.
(4) The first speaker (A) counters with an additional proposition, and the second speaker (B) agrees with the first speaker's proposition.
(5) 2–4 is repeated one more time with other propositions.
(6) The second speaker (B) finally accepts the first speaker's (A's) stance.

The length of a dialogue is fixed with 27 utterances: 15 utterances are fixed phrases such as "You have a point" and 12 utterances correspond to propositions rewritten into utterances by the methods except for Vanilla. An example of the original scenario generated from the argumentation structures is shown in Table 1. The propositions will be rewritten into utterances and the other utterances are used as they are for creating the evaluated dialogues.

Flow	ID	Prop	Proposition or Fixed Utterance
1	U_1	✓	B: Are you for vs. against autonomous cars?
	U_2	✓	A: Autonomous cars are good.
	U_3		B: You have a point.
	U_4	✓	B: Autonomous cars are bad.
2	U_5		A: Hmmm...
	U_6	✓	A: If autonomous cars are realized, there will be fewer traffic accidents.
3	U_7		B: Hmmm...
	U_8	✓	B: Autonomous cars controlled by artificial intelligence are unreliable.
4	U_9		A: You have a point.
	U_{10}	✓	A: Autonomous cars can prevent accidents involving drunk drivers.
	U_{11}		B: Indeed, that may be true.
	U_{12}		A: In other words,
	U_{13}	✓	A: If autonomous cars are realized, there will be fewer traffic accidents.
	U_{14}		B: Certainly, that may be true.
...	...	...	...

Table 1: Example of original scenario generated from argumentation structure. 'Flow' column corresponds to numbers in the argumentation flow described in Section 5.2. 'Prop' (proposition) column's check indicates that the utterance is from a proposition and will be rewritten into utterances by the proposed method.

For the evaluation, we created ten dialogue scenarios with randomly selected propositions for the five argumentation themes and the four methods, resulting in 200 dialogues in total (10 dialogue scenarios $\times$ 5 themes $\times$ 4 generation methods $=$ 200 dialogues). Each method except for Vanilla rewrote the propositions in the 50 dialogue scenarios and generated 50 dialogues for the evaluation. Note that the 200 dialogues automatically created from the argumentation structures for the evaluation are not related to the 250 dialogues in the argumentative dialogue corpus because those are only used for fine-tuning the language model.

5.3 Evaluation Procedure

We prepared three metrics for evaluating the quality of generated argumentative dialogues. We used a seven-point Likert scale (1: strongly disagree, 7: strongly agree) according to the degree of agreement with the following statements.

(1) **Grammar:** Grammar is appropriate.
(2) **Naturalness:** The contents and phrases in each utterance naturally reflect the previous context.

[1] https://huggingface.co/sentence-transformers

(3) **Persuasiveness:** The dialogue is persuasive in terms of consistency throughout the dialogue.

Five crowdworkers were recruited through a Japanese crowdsourcing platform[2]. They were instructed to judge each metric independently. Each crowdworker evaluated 200 shuffled dialogues.

5.4 Results and Discussion

Table 2 shows the results of manually evaluating the generated dialogues from the four methods. The proposed method performs the best in terms of all the metrics and has significantly better naturalness than the other methods (two-tailed binomial test, Bonferroni corrected $p < 0.05$). Since the proposed method is evaluated as equally persuasive as Vanilla, consistency was maintained when rewriting the proposition. We assume that the persuasiveness was not improved from Vanilla because the content of each proposition is the same as that of the original in the argumentation structure. The persuasiveness of Ret-Rewrite was low probably due to the difficulty of retrieving an appropriate proposition from an utterance in creating the training data for fine-tuning; for example, an irrelevant proposition tended to be retrieved, thus leading to an inappropriate rewrite.

Figure 3 shows the examples of generated dialogues from the four comparative methods. Ret-Rewrite and Kwd-Rewrite generated erroneous utterances such as speaker B's first utterance in Kwd-Rewrite ("I disagree with autonomous cars, but I agree with them"). In Kwd-Rewrite, B's third utterance ("Autonomous cars are bad" in the proposition) was incorrectly rewritten into a question without mentioning B's stance ("Are autonomous cars good?"). The proposed method successfully generated a dialogue with phrases such as "I am for" and "I think," resulting in natural dialogue.

Method	Grammar	Naturalness	Persuasiveness
(a) Vanilla	4.42	4.49_{b}	3.90_{b}
(b) Ret-Rewrite	4.54	3.30	2.33
(c) Kwd-Rewrite	4.31	4.40_{b}	3.62_{b}
(d) Proposed	$\mathbf{4.68}_{c}$	$\mathbf{4.76}_{abc}$	$\mathbf{3.96}_{b}$

Table 2: Average scores over judges for the generated dialogues. Subscripts indicate significant difference from corresponding models.

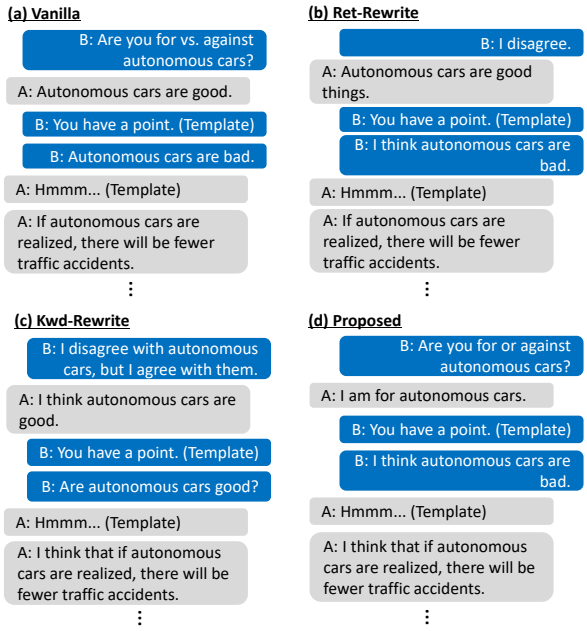

Figure 3: Sample of generated argumentative dialogue (first six utterances) from four methods. 'Template' indicates that the utterance is created by a hand-crafted template.

6 Conclusion

This study proposed a method to generate natural argumentative dialogue by combining an argumentation structure and language model. We proposed the method to fine-tune the language model to rewrite propositions of an argumentation structure into a natural argumentative dialogue on the basis of their key information, such as keywords and stance, into the next utterance while considering its context. The proposed method significantly improved the naturalness of dialogues without losing the consistency of argumentation.

Future work includes conducting a live evaluation of the proposed method and validating the effectiveness of the proposed method in other experimental settings using other pre-trained language models and datasets. In addition, the naturalness of generated dialogues needs to be improved by automatically generating more naturalistic dialogue scenarios by using a language model.

Acknowledgments

This study was conducted in connection with our joint research project with Osaka University and Advanced Telecommunications Research Institute International (ATR). We would like to acknowledge Dr. Kazuki Sakai, Prof. Yuichiro Yoshikawa, and Dr. Takashi Minato for their helpful discussions and comments.

[2] https://www.lancers.jp

References

Christopher Hidey and Kathleen McKeown. 2019. Fixed that for you: Generating contrasive claims with semantic edits. In *Proc. of NAACL-HLT*, pages 1756–1767.

Ryuichiro Higashinaka, Kazuki Sakai, Hiroaki Sugiyama, Hiromi Narimatsu, Tsunehiro Arimoto, Takaaki Fukutomi, KiyoakiMatsui, Yusuke Ijima, Hiroaki Ito, Shoko Araki, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Yoshihiro Matsuo. 2017. Argumentative dialogue system based on argumentation structures. In *Proc. of SEMDIAL*, pages 154–155.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proc. of ACL*, pages 219–230.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.0128*, pages 1–6.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, pages 1–11.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*, pages 1–11.

Koh Mitsuda, Ryuichiro Higashinaka, Taichi Katayama, and Junji Tomita. 2019. Generating supportive utterances for open-domain argumentative dialogue systems. In *Proc. of IWSDS*, pages 1–12.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2015. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In *Proc. of PACLIC*, pages 307–314.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proc. of COLING*, pages 3349–3358.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

ChaeHun Park, Wonsuk Yang, and Jong C. Park. 2019. ArgDiver: Generating sentential arguments from diverse perspectives on controversial topic. In *Proc. of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 56–65.

Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2018. Utilizing argument mining techniques for argumentative dialogue systems. In *Proc. of IWSDS*, pages 1–12.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, pages 1–67.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn Walker. 2017. Debbie, the debate bot of the future. In *Proc. of IWSDS*, pages 1–6.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, pages 1–11.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*, pages 1–10.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, pages 1–25.

Kazuki Sakai, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Junji Tomita. 2020. Hierarchical argumentation structure for persuasive argumentative dialogue generation. *IEICE Transactions on Information and Systems*, E103.D(2):424–434.

Kazuki Sakai, Akari Inago, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Junji Tomita. 2018. Creating large-scale argumentation structures for dialogue systems. In *Proc. of LREC*, pages 3975–3980.

Misa Sato, Kohsuke Yanai, Toshihiko Yanase, Toshinori Miyoshi, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proc. of ACL-IJCNLP System Demonstrations*, pages 109–114.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based Japanese chit-chat systems. *arXiv preprint arXiv:2109.05217*, pages 1–11.

Stephen Edelston Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Douglas Walton. 2013. *Methods of argumentation*. Cambridge University Press.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proc. of The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proc. of AAAI*, pages 3351–3357.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proc. of EMNLP*, pages 2831–2845.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proc. of AAAI*, pages 730–738.

Qingfu Zhu, Weinan Zhang, Lei Cui, and Ting Liu. 2019. Order-sensitive keywords based response generation in open-domain conversational systems. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(2):1–18.