DiSCoL: Toward Engaging *Di*alogue *S*ystems through *Co*nversational *L*ine Guided Response Generation

Sarik Ghazarian,[1] **Zixi Liu,**[1] **Tuhin Chakrabarty,**[2]
Xuezhe Ma,[1] **Aram Galstyan,**[1] **Nanyun Peng**[1, 3]
[1] University of Southern California / Information Sciences Institute
[2] Computer Science Department of Columbia University
[3] Computer Science Department of University of California, Los Angeles
{sarik, zixiliu, xuezhema, galstyan}@isi.edu
tuhin.chakr@cs.columbia.edu violetpeng@cs.ucla.edu

Abstract

Having engaging and informative conversations with users is the utmost goal for open-domain conversational systems. Recent advances in transformer-based language models and their applications to dialogue systems have succeeded in generating fluent and human-like responses. However, those systems still lack control over the generation process toward producing contentful responses and achieving engaging conversations. To address this, we present **DiSCoL** (**Di**alogue **S**ystems through **Co**versational **L**ine guided response generation). DiSCoL is an open-domain dialogue system that leverages conversational lines (briefly **convlines**) as controllable and informative content-planning elements to guide the generation model in producing engaging and informative responses. Two primary modules in DiSCoL's pipeline are conditional generators trained for 1) predicting relevant and informative convlines for dialogue contexts and 2) generating high-quality responses conditioned on the predicted convlines. Users can also change the returned convlines to *control* the direction of the conversations toward topics that are more interesting for them. Through automatic and human evaluations, we demonstrate the efficiency of the convlines in producing engaging conversations.

1 Introduction

Over the past decade, users have actively engaged with dialogue systems to fulfill a wide range of requirements. *Task-oriented dialogue systems* have assisted users in accomplishing specific tasks such as finding apartments (Gustafson et al., 2000) and restaurants (Gruenstein and Seneff, 2007) or even booking movie tickets (Li et al., 2017). While, *Open-domain dialogue systems* have been extensively leveraged for psychotherapy counseling, entertainment, and even teaching foreign languages to users (Zhou et al., 2020; Oh et al., 2017; Sarosa

Figure 1: A dialogue context and its three responses generated based on DialoGPT and our proposed DiS-CoL system using originally inferred and manipulated convlines, respectively. DiSCoL leverages convlines (depicted in colored boxes) to guide the generation model to encapsulate those informative contents. Our demo enables the user to edit or remove the inferred convlines (shown in blue for edits and red for removal) to guide the conversation towards its desired directions.

et al., 2020). In this work, we focus on the second group.

In the context of open-domain dialogue systems, neural-network-based generative models have outperformed retrieval-based systems by generating diverse and novel responses. More recently, large-scale language models with transformer-based architectures, such as GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2019), have advanced the state of the art in Natural Language Generation and Dialogue Systems. Such models can be further enhanced by fine-tuning them on task-specific data, as it is the case of DialoGPT (dialogue generative pre-trained transformer) (Zhang et al., 2019), a neural conversational response generation model, trained on 147M conversation-like exchanges extracted from Reddit. Although responses generated by such models are fluent and locally coherent, they usually suffer from content poverty (e.g., generating non-informative content), which can negatively impact user engagement. Furthermore, these models do not allow the users to exert control on the generation process and guide the conversation to-
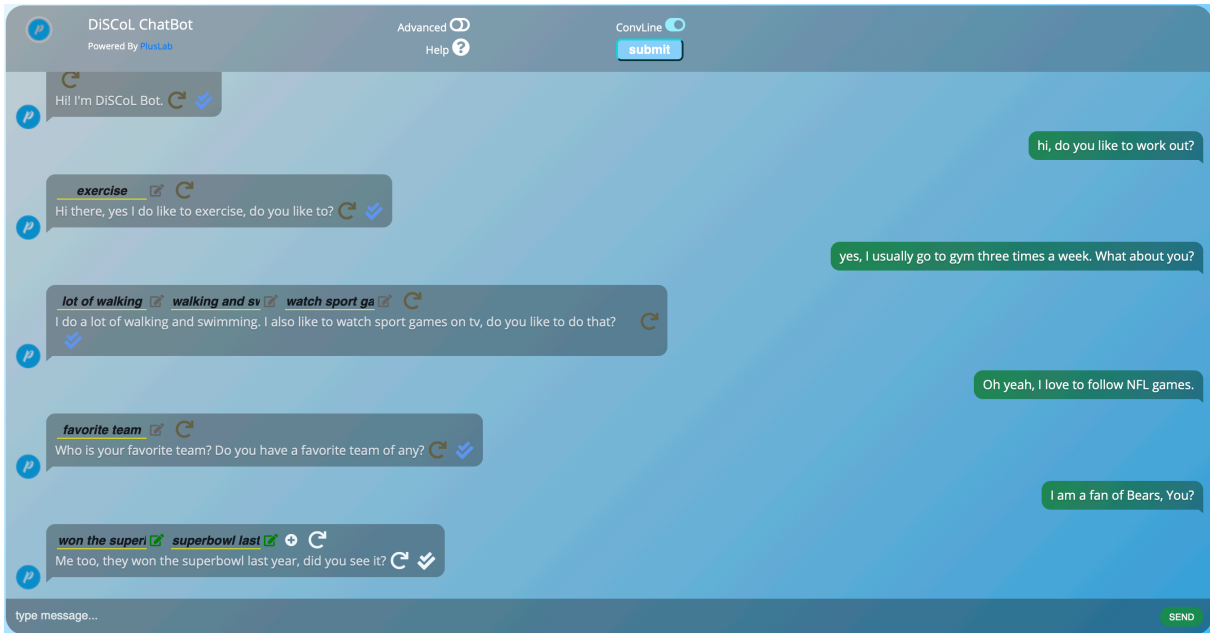
26

Figure 2: A snapshot of the proposed DiSCoL system

ward users' desired direction. The first block in Figure 1 depicts an example of a generated response by DialoGPT.

To alleviate this issue, here we propose DiSCoL, an open-domain dialogue system, which leverages **convlines** as primary elements to add *control* for generating informative and content-rich responses. Convlines are abstract representations of utterances in the dialogues that can be used as content planning elements to form high-level content of an utterance and guide the generator to incorporate these informative units in the generation (See colored boxes in Figure 1). Content planning has been shown to be beneficial in the story generation task. These abstract representations known as storylines or story plots have been successful to guide the language models produce more coherent and fluent stories (Yao et al., 2019; Goldfarb-Tarrant et al., 2019; Fan et al., 2019; Goldfarb-Tarrant et al., 2020; Rashkin et al., 2020; Brahman et al., 2020).

DiSCoL is composed of four main neural-network-based modules (See Figure 3). The first two modules are designed to extract entities and topics of the dialogue context. The third module is a fine-tuned conditional generator that learns to take the dialogue context and previously extracted information and predict convlines that would be leveraged in the response generator module. Similar to convline generator, response generator is a conditional auto-regressive language model that generates response conditioned on the dialogue context and its convlines, entities, and topics ex-

tracted from previous modules. The middle block of Figure 1 exhibits the generated response for the inferred convlines shown in green boxes. In the interactive setting of our devised demo from which a snapshot is shown in Figure 2, we provide the users with the facility to manipulate the predicted convlines to direct the conversation toward its topics of interest. The last block in Figure 1 depicts the removed and edited convlines (red and blue boxes) that led the generator to generate a slightly different response by taking into account the applied adjustments.

We validate DiSCoL on the Topical chat dataset (Gopalakrishnan et al., 2019) using both human and automatic evaluations. Our results demonstrate the superiority of DiSCoL over DialoGPT in terms of generating higher quality responses, thus indicating the usefulness of convlines as dialogue control mechanisms for generating more engaging responses. We release the source code and trained models to facilitate the future dialogue research. [1]

2 system Architecture

The architecture of our proposed DiSCoL demo system and its modules are depicted in Figure 3. A user converses with the system by writing an utterance as an input. This utterance passes through all the modules and in each module some new information such as its extracted entities, topics, and convlines are augmented. The last module, response

---

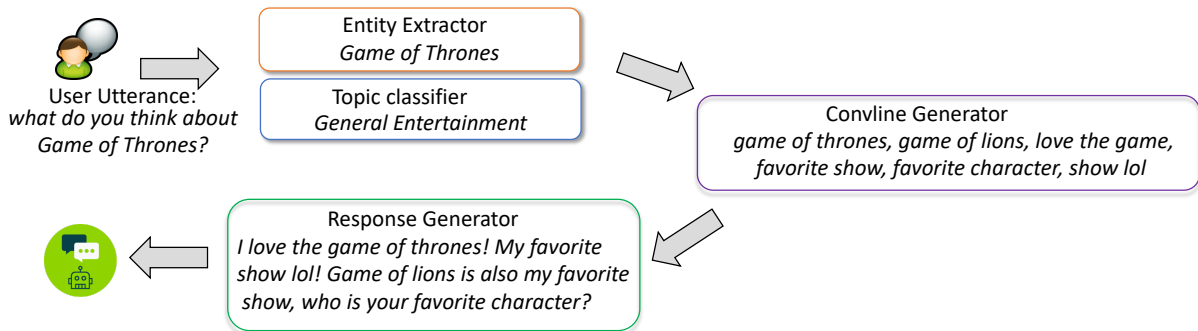[1]Github Link: https://github.com/PlusLabNLP/Dialogue_System_Hackathon

Figure 3: Architecture of DiSCoL system

generator, incorporates all this information to generate a response as the output of the system. In this section, we explain each module in detail.

2.1 Entity Extractor

One of the principal components in the conversational systems is the set of entities that both interlocutors are interested to converse about. It is crucial that the system can identify the main entities from the dialogue context and try to continue the conversation by providing more relevant information or even expressing its opinions and impressions regarding them. Therefore, in DiSCoL we take the user's utterance as the dialogue context and extract its entities. This task is known as a named entity recognition (NER) task, where each token in the text is classified into one of the predefined classes such as a person, organization, location or other.

Toward this goal, we leverage the BERT model (Devlin et al., 2019) fine-tuned on CoNLL-2003 dataset (Sang and De Meulder, 2003), which is a well-known corpus for NER task.[2] We detokenize the output of the fine-tuned BERT model to get the original version of entities' tokens and disregard the predefined classes of entities since in our case they do not augment additional benefits. As shown in Figure 3, all entities with labels other than *O* are returned from the entity extractor module.

2.2 Topic Classifier

Knowing the topic that the user is enthusiastic to discuss is essential for the dialogue system to generate utterances about that specific topic. The blue box in Figure 3 represents the topic classifier that takes the user's utterance and predicts the most relevant topics from a predefined set. These topics

are later used for predicting convlines and consequently generating responses.

Due to the proven effectiveness of the BERT model (Devlin et al., 2019) and its wide applicability in many classification tasks, we incorporate it into the topic classifier module of DiSCoL. We fine-tune BERT model on pairs of utterances and their aligned topics with the main goal of minimizing the cross-entropy loss.

2.3 Convline Generator

DiSCoL's main contribution is in the convline generator module that is depicted as the purple box in Figure 3. Convlines are abstract representations or content plans of utterances throughout the conversation. These representations, which are also known as storylines or story plots in the context of story generation, have recently posited their efficiency in generating higher quality stories (Yao et al., 2019; Fan et al., 2019; Goldfarb-Tarrant et al., 2020; Rashkin et al., 2020). Story generation models leverage plan-and-write framework that is successful in generating fluent and informative stories by the intervention of storylines as an intermediate step. In this work, we follow the same idea but in the context of conversational systems. In particular, we aim to show that the controlled generation of high-quality utterances by planning in advance and leveraging useful abstract-level convlines can be beneficial for dialogue systems as well.

To compose the convlines as the main component in the convline generator module, we extract sequences of important words in each utterance from existing human-human conversational data. We use the YAKE (Campos et al., 2018) method that relies on the text's statistical features to extract the most important keywords of an utterance, as it has shown its superiority over other state-of-the-art unsupervised approaches such as TF-IDF and RAKE (Rose et al., 2010).

---

[2]We leverage fine-tuned BERT model provided by Huggingface (`https://github.com/huggingface/transformers`).

28

To train the convline generator, we extract pairs of (u_i, r_i) as a set of consecutive pairs of dialogue context utterances and their corresponding ground-truth responses in the human-human conversational data. For each dialogue context utterance (u_i), we extract its entities (e_i) and topics (t_i) using the entity extractor and topic classifier modules. Each response (r_i) is replaced by its convlines (c_i) obtained by the YAKE algorithm. The constructed input data are in (u_i, e_i, t_i, c_i) format.

The convline generator is a conditional model that generates the most probable convlines given the provided dialogue context utterance together with its entities and topics. To this end, we apply BART (Lewis et al., 2019), which is a state-of-the-art pre-trained sequence-to-sequence generative model. It combines a bidirectional encoder as that of BERT (Devlin et al., 2019) to encode the input and a GPT like (Radford et al., 2018) auto-regressive decoder model to generate convlines as the output. The top block in Figure 4 encapsulates the training process of the convlines module. We fine-tune BART on the constructed training data with the objective of minimizing the negative log likelihood shown in Equation (1).

$$L_{line\_gen} = -log \sum_{i=1}^{n} P(c_i | u_i, t_i, e_i) \quad (1)$$

During inference, the fine-tuned BART model takes the user's utterance augmented with its inferred entities and topics to predict the most probable convlines, as depicted in the bottom block of Figure 4. We use top-k sampling (Fan et al., 2019) with $k = 5$ and a temperature of 0.7 for the generation.

2.4 Response Generator

The last module in DiSCoL system's pipeline is the response generator that is identical to convline generator except for the type of inputs and outputs. The response generator takes the dialogue context utterance, its convlines and topics as inputs and generates response conditioned on those data.

$$L_{resp\_gen} = -log \sum_{i=1}^{n} P(r_i | u_i, t_i, c_i) \quad (2)$$

During training, we provide utterances, their topics and convlines extracted from YAKE to the BART model and fine-tune this pre-trained conditional generator. As it is shown in Equation (2), the training objective is to maximize the probability of generating ground-truth responses given their context utterances, topics, and the convlines.

During inference, the generator attempts to produce the most probable responses that include convlines returned by the convline generator module.

3   System Implementation

We test our system on Topical-Chat dataset (Gopalakrishnan et al., 2019) that includes knowledge-grounded human-human conversations covering a set of 8 different topics. This dataset has been collected by employing Amazon Mechanical Turk (AMT) workers who have been provided with specific entities and some external knowledge (Wikipedia lead sections, Washington Post articles, or some Reddit fun facts) to chat about. Therefore, each utterance in the conversation is either based on provided knowledge sources or the user's personal knowledge. Overall, 261 popular entities spanning 8 various topics (Fashion, Sports, Books, Politics, General Entertainment, Music, Science & Technology and Movies) have been selected for the dataset collection. We add *General* topic for utterances (e.g. greetings) that do not include any specific contents such as *"hi, how are you today?"*.

3.1   Topic Classification Data

Although each utterance in the Topical-Chat dataset comes from either provided external knowledge or interlocutor's personal knowledge about some specified entities, it lacks determined topic labels, which are necessary for DiSCoL modules. To infer topics, we first manually match all 261 entities in the external knowledge to one of the topics in the predefined set (Fashion, Sports, Books, Politics, and etc.). Next, we label all utterances talking about those entities to their corresponding topics. This simple labeling scheme produces topics for about 78% of the 188,378 *(easy_set)* total utterances. As an example, the utterance *"Do you know Tom Brady"* is about *"Tom Brady"* entity that is an indication of the *"Sports"* topic. Therefore, we label this utterance with the *"Sports"* topic.

The remaining challenging utterances are mainly the continuation of the dialogue history without directly containing any entities. Take *"I guess they live up to their name then!"* as an example of such utterances with no mentioned entities. We pursue the following context-based heuristics to label such *challenging_set* utterances with their relevant topics. If the utterance's neighbors (utterances right before or after the current utterance) are from *easy_set* and both share the same entity, we assign
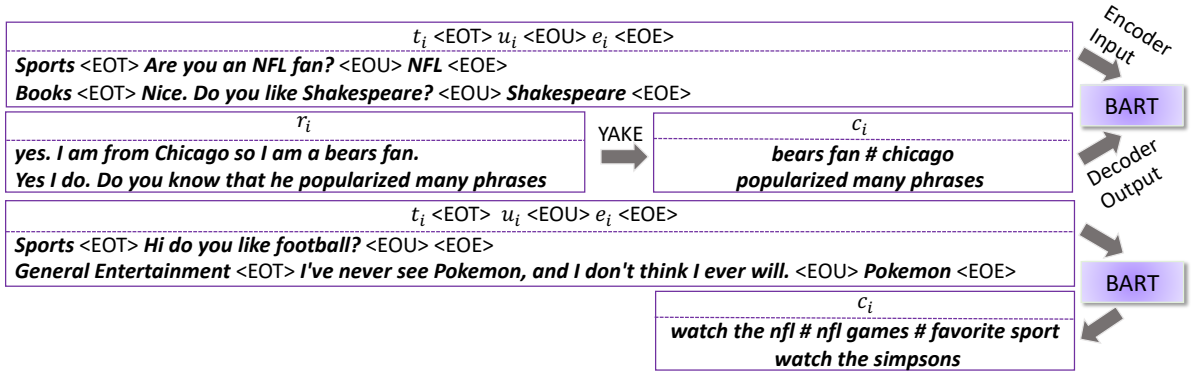
Figure 4: Architecture of the convline generator during training and inference time

Uttr.	Easy_set	Challenging_set	General_uttr.
188,378	146,370	5,323	5,966

Table 1: Statistics of different groups of utterances (uttr.) in the Topical chat dataset

Dialogue Context	Annotators	Kappa	Pearson
100	33	0.44	0.5

Table 2: Statistics and inter-annotator agreements of AMT evaluations on DiSCoL and DialoGPT performances.

that entity's topic to the current utterance, while in the case of neighbors containing different entities, we label the given utterance with both utterances' topics. If the previous rules do not apply to an utterance in the *challenging_set*, we use the most frequent topic in the dialog as its topic.

In parallel to the above heuristics and in order to improve the quality of assigned topics, we also apply a keyword-based classifier that classifies *challenging_set* utterances with appropriate topics. The keyword-based classifier retrieves the most similar entity from the overall 261 entities to each utterance's keywords using their BERT embeddings. Then, the manually matched topics for the retrieved entity are assigned to the utterance. We only consider 5323 *challenging_set* utterances that their adapted labels based on both approaches: 1) context-based heuristics and 2) keyword-based classifier are the same (See statistics in Table 1).

The remaining utterances shown in the last column of Table 1 are mainly general utterances for starting or ending conversations without any specific content such as *"Good Morning! How are you today?"* or *"It was nice chatting with you!"*. We fine-tune the BERT model as the topic classifier for 10 epochs and get an accuracy of 85.55 on the validation set.

3.2 Convline Generator Data

Convlines are the central components in the training of the DiSCoL system. We leverage YAKE (Campos et al., 2018) for retrieving discourse keywords representing convlines. YAKE assigns an importance score to tokens in a text by following an unsupervised approach that builds upon features extracted from the text (Campos et al., 2018). In this model, a set of features are computed for each term in the text. Subsequently, a list of candidates (n-grams of tokens) is created. Next, the Levenshtein distance is used to remove duplicate keywords. Finally, the aggregation of token scores in each keyword is used to represent the keyword's score. Keywords with lower scores are returned as the text's salient convlines. We use YAKE to generate a contiguous sequence of 1, 2, and 3-grams candidate convlines. We extract 3-grams convlines, followed by extracting 2-grams and 1-gram that are not included in the previously returned keywords. We fine-tune BART-large for both convlines and response generator models for 3 epochs and checkpoint the best epoch based on validation perplexity.[3]

4 Experimental Results

We evaluate the performance of DiSCoL system against DialoGPT, which is one of the strongest recent baselines that has shown its efficiency in generating consistent and relevant responses.

4.1 Metrics

To explore the efficiency of our proposed controlled response generation, we apply both automatic and human evaluations.

---

[3]We fine-tune BART model using https://github.com/pytorch/fairseq

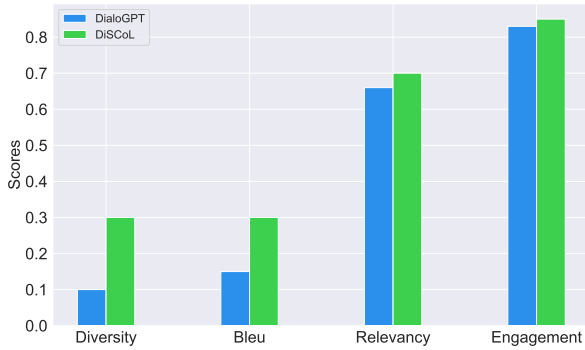Figure 5: Automatic evaluations on responses generated by DiSCoL and DialoGPT systems

Figure 6: Human evaluations on responses generated by DiSCoL and DialoGPT systems

4.2 Automatic Evaluations

Due to the multi-faceted nature of dialogue quality, it is necessary to do the evaluation from different aspects (See et al., 2019; Mehri and Eskenazi, 2020). To this end, we compare the quality of DiSCoL and DialoGPT generated responses through computing different metrics. We conduct automatic evaluations and compute evaluation metrics on 23,530 consecutive utterance pairs (dialogue context utterances and their ground-truth responses) of the Topical chat test set. The measured metrics are averaged over all utterance pairs within the test set. We compute BLEU-3 (Papineni et al., 2002) to evaluate the similarity of generated responses to ground-truth responses based on the 3-grams overlaps. Due to the one-to-many essence of open-domain dialogue systems and the imperfection of such word-overlap metrics (Liu et al., 2016; Ghazarian et al., 2019; Mehri and Eskenazi, 2020), we also focus on three main aspects: diversity, relevancy, and engagingness as better indications of systems performances.

Diversity measures the percentage of distinct generated tokens by each model. Li et al. (2015) proposed distinct-2 that computes distinct bi-grams divided by the total number of generated words. Relevancy utilizes both dialogue context utterance and the generated response to deliberate how much it is relevant to the given utterance (Tao et al., 2018; Ghazarian et al., 2019). We use the contextualized Ruber metric for this purpose (Ghazarian et al., 2019). At the end, since in open-domain dialogue systems, it is necessary to have both relevant and interesting responses to make the user feel satisfied (Ghazarian et al., 2020), we further validate systems based on the engagingness of responses. We compute engagingness as the probability score of the engaging class predicted by Ghazarian et al. (2020)'s proposed engagement classifier.
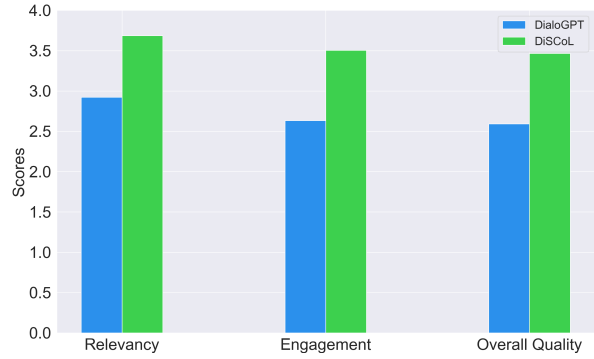
4.3 Human Evaluations

We extend our evaluations by running AMT experiments to report human judgments on the quality of system-generated responses. We randomly select 100 dialogue context utterances from the Topical chat test set. For each given dialogue context utterance, we ask three AMT workers to rate DiSCoL and DialoGPT's generated responses by keeping these systems anonymous. Participants rate the relevancy, engagingness, and overall quality of each response on a 5-point Likert scale (1 indicating irrelevant/not engaging and low-quality response). The statistics of the AMT experiment are shown in Table 2.

4.4 Results

Automatic Evaluation. Figure 5 depicts the average scores of diversity, BLEU, relevancy, and engagingness resulted from automatic evaluation metrics for all the generated responses of DiSCoL and DialoGPT systems. The strength of DiSCoL is noticeable from its higher BLEU score and more diverse, relevant, and engaging responses. Overall, the diversity is low due to the limited distinct topics considered in the Topical chat dataset. The BLEU metric is low for both systems which shows its inadequacy in the open-domain evaluations; where a response can be super appropriate and at the same time not similar to the ground-truth response.

Human Evaluation. The bars in Figure 6 demonstrate the average of human annotations for different qualities of generated utterances. Each response's score is the mean aggregation of three annotators' ratings. According to Figure 6, annotators appraise responses generated by DiSCoL with higher scores in terms of relevancy, engagingness, and overall quality. This could be an evidence for the positive impact of incorporating convlines to

guide the dialogue system towards generating controllable, relevant, and contentful responses that infuse the user to converse for a longer time.

5 Conclusion

We have introduced DiSCoL, an open domain dialogue system that leverages convline as an intermediate step toward generating more informative and controllable responses in dialogues. The convlines are predicted and subsequently leveraged in the response generation process. Additionally, DiSCoL allows users to manipulate convlines towards their favorite conversational direction. Our findings show that in contrast to other transformer-based dialogue systems that do not incorporate content planning, DiSCoL takes the advantage of such a principled structure to generate better and more engaging conversations with users.

In the future, we imagine an open path of possible research in the controllable conversations which would guide the dialogue toward having pleasant features such as empathy and bias-free or even personalized convlines to generate dialogues with such aspects. It is also expecting to train dialogue models to converse by following specific styles such as generating formal conversations by predicting more formal convlines.

6 Ethics

Through the entire phases of the conducted research and developed DiSCoL system, all co-authors were agreed and adhered to *ACM Code of Ethics*. Our effort was to ensure we stuck to the conscience of the profession and considered the Code principles. We certify that this system and all the presented evaluations are compatible with the provided code. In the following, we discuss two main spots in the development and evaluation of our system that could be targeted for encompassing abusive and improper conversations and having biased evaluations.

DiSCoL System's Development   The main contribution of our proposed DiSCoL system is to augment controllable response generation with the intervention of convlines that leads the generation towards producing more relevant and interesting responses. Indeed, DiSCoL provides an opportunity for users to manipulate the convlines and guide the system to continue the conversation in the user's favorite direction. All DiSCoL's modules

leverage pre-trained large language models such as BART (Lewis et al., 2019) and fine-tune them on recently proposed Topical chat dataset (Gopalakrishnan et al., 2019). One potential harm that DiSCoL could cause is its feasibility to generate improper responses conditioned on the inferred convlines with abusive contents. Since the convline and response generators are BART models finetuned on human-human conversations that do not encompass profanity and inappropriate content ((Gopalakrishnan et al., 2019)), hence the convlines that indeed are important informative units of the utterances would be free of bias and obscene content. However, there still is a possibility of dual-usage attacks by augmenting conversations with offensive languages to fine-tune the generators and teach them to generate such inappropriate content. The identification of such attacks that could occur in almost all learnable models and the way to overcome them by itself is a distinct and huge research area that is out of this paper's scope.

DiSCoL System's Evaluation   Alongside the automatic evaluation for demonstrating the efficiency of controllable generations using convlines, we further collected human annotations by conducting Amazon Mechanical Turk (AMT) experiments. We provided different systems responses for given utterances while keeping systems anonymous and asked users to rate responses by considering different aspects that had been explained in the AMT surveys. We estimated the average time users would spend on each survey and fairly compensated them according to the hourly wage.

We kept the privacy of all AMT turkers who participated in the experiments. Our experiments did not have the requisite to know the user's personal information, therefore their personal information including their genre, ethnicity, and etc. are not revealed. This fades the necessity for IRB approvals.

At the end, we want to note that our system's target is NLP open-domain conversational AI community with the main goal of achieve engaging conversations with the incorporation of convlines and increasing the user's ability to control the generation process. Likewise other proposed dialogue systems, we anticipate specific failure modes specifically for novel conversations on new topics. Lifelong learning in dialogue systems which is not the focus of this work is a research area that attempts to enhance conversation systems' ability to deal with such novel scenarios.

Acknowledgment

References

Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. 2020. Cue me in: Content-inducing approaches to interactive story generation. In *Asia-Pacific Chapter of the Association for Computational Linguistics (AACL)*.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pages 806–810. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Association for Computational Linguistics (ACL)*.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen workshop of NAACL-HLT)*.

Sarik Ghazarian, Ralph M Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *AAAI*, pages 7789–7796.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019), Demonstrations Track*, volume 4, pages 89–97.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*.

Alexander Gruenstein and Stephanie Seneff. 2007. Releasing a multimodal dialogue system into the wild: User support mechanisms. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 111–119.

Joakim Gustafson, Linda Bell, Jonas Beskow, Johan Boye, Rolf Carlson, Jens Edlund, Björn Granström, David House, and Mats Wirén. 2000. Adapt—a multimodal conversational dialogue system in an apartment domain. In *The Sixth International Conference on Spoken Language Processing (ICSLP), Beijing, China*, pages 134–137.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Association for Computational Linguistics (ACL)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *International Joint Conference on Natural Language Processing (IJCNLP)*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. 2017. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 371–375. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

M Sarosa, M Kusumawardani, A Suyono, and MH Wijaya. 2020. Developing a social media-based chatbot for english learning. In *IOP Conference Series: Materials Science and Engineering*, page 012074. IOP Publishing.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.