Learning to Plan and Realize Separately for Open-Ended Dialogue Systems

Sashank Santhanam[1], Zhuo Cheng[1], Brodie Mather[2], Bonnie Dorr[2],
Archna Bhatia[2], Bryanna Hebenstreit[2], Alan Zemel[3],
Adam Dalton[2], Tomek Strzalkowski[4] and Samira Shaikh[1]

University of North Carolina, Charlotte, NC, USA[1]
Institute for Human and Machine Cognition (IHMC), Ocala, FL, USA[2]
State University of New York, Albany, NY, USA[3]
Rensselaer Polytechnic Institute, Troy, NY, USA[4]
{ssantha1, samirashaikh}@uncc.edu[1]

Abstract

Achieving true human-like ability to conduct a conversation remains an elusive goal for open-ended dialogue systems. We posit this is because extant approaches towards natural language generation (NLG) are typically construed as end-to-end architectures that do not adequately model human generation processes. To investigate, we decouple generation into two separate phases: planning and realization. In the planning phase, we train two planners to generate plans for response utterances. The realization phase uses response plans to produce an appropriate response. Through rigorous evaluations, both automated and human, we demonstrate that decoupling the process into planning and realization performs better than an end-to-end approach.

1 Introduction

Recent advancements in the area of generative modeling have helped increase the fluency of generative models. However, several issues persist: coherence of output and the semblance of mere repetition/hallucination of tokens from the training data (Moryossef et al., 2019; Wiseman et al., 2017). One reason could be that the generation task is typically construed as an end-to-end system. This is in contrast to traditional approaches, which incorporate a sequence of steps in the NLG system, including content determination, sentence planning, and surface realization (Reiter, 1994; Reiter and Dale, 2000). A review of literature from psycholinguistics and cognitive science also provides strong empirical evidence that the human language production process is not a monolith (Dell, 1985; Bock, 1996; Bock et al., 2007; Kennison, 2018).

Prior approaches have indeed incorporated content planning into the NLG system, for example data-to-text generation problems (Puduppully et al., 2019; Moryossef et al., 2019) as well as classic
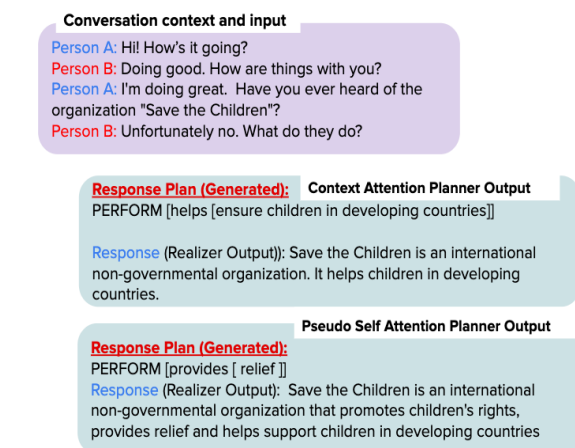

Figure 1: Example conversation between two speakers A & B where the response for the speaker B is generated based on the response plan from two learned planners: Context Attention and Pseudo Self Attention.

works that include planning, based on speech acts (Cohen and Perrault, 1979) (for an in-depth review c.f. (Garoufi, 2014)). Our work closely follows these prior approaches, with one crucial difference: our planners are not based on dialogue acts or speech acts.

Consider the example in Fig. 1. An input utterance by Person B, a statement (*Unfortunately no.*), followed by a question (*What do they do?*), can be effectively responded to using plans, learned and generated, prior to the realization phase. The realization output can then include the mention of *provides relief*, consistent with the generated plan (*PERFORM [provides [relief]]*).

Dialogue acts (Stolcke et al., 2000) (e.g., statements, questions), by their nature, encompass a wide variety of realized output, and hence cannot sufficiently constrain the language model during the generation process. Research has addressed this issue by adapting existing taxonomies (Stolcke et al., 2000) towards their own goals (Wu et al.,

2018; Oraby et al., 2017). We instead use an adapted and extended form of lexical-conceptual structures (LCSs) to help constrain the realization output more effectively (Dorr, 1994).

Our work makes the following contributions:
• We investigate the impact of separating planning and realization in open-domain dialogue and find that the approach produces better responses per automated metrics and detailed human evaluations.
• We propose the use of LCS-inspired representations based on asks and framings, which in turn are grounded in conversation analysis literature, to generate plans, instead of using dialogue acts.
• We release corpora annotated with plans for all utterances, using three planners, including symbolic planners and attention-based planners.

2 Related Work

Open-Ended Dialogue Systems:

Transformer models (Vaswani et al., 2017) and large transformer-based language models such as GPT, GPT-2, XLNet, BERT (Radford et al., 2018, 2019; Yang et al., 2019; Devlin et al., 2019) have helped achieve the SOTA performance across several natural language tasks. However, these models do not achieve the same level of consistent performance on generative modeling tasks as opposed to language understanding tasks (Ziegler et al., 2019; Edunov et al., 2019). Wolf *et al.* (2019) propose a transfer learning approach that fine tunes large pretrained language models and achieves SOTA scores on the PERSONA-chat dataset (Golovanov et al., 2019) and in the CONVAI2 competition (Dinan et al., 2019; Yusupov and Kuratov, 2018). Keskar *et al.* (2019) introduce a large-scale conditional transformer model that improves generation based on control codes.

Our training paradigm is consistent with existing research that constrains large-scale language models across generation tasks (Rashkin et al., 2019; Urbanek et al., 2019) and yields controllable text generation (Shen et al., 2019; Zhou et al., 2017), with one key difference: we learn to plan and realize separately. Accordingly, we overview planning based approaches next.

Planning-Based Approaches: A standard component of traditional NLG systems is a planner (Reiter and Dale, 2000). Prior work leverages intent and meaning representations (MR) to understand the content of the message (Young et al., 2013), but largely in task-oriented as opposed to open-ended

dialogue systems (He et al., 2018). Novikova *et al.* (2017) propose the E2E challenge and use MRs to show lexical richness and syntactic variation. Similarly, Gardent *et al.* (2017) focus on structured data (e.g. DBpedia) to generate text in the WebNLG framework. Moryossef *et al.* (2019) use an explicit symbolic component for planning in a neural data to text generation system that allows controllable generation. Along with conversational intents, dialogue acts are also used for natural language understanding (NLU) in task-oriented systems (Li et al., 2019; Peskov et al., 2019).

In contrast to these prior approaches, our work uses more in-depth meaning representations for open-domain dialogue systems based on lexical conceptual structures (explained in Section 3.1).

3 Approach

3.1 NLU using Asks and Framing

The representation we use to generate plans leverages *asks* and *framings* based on conversation analysis literature (Pomerantz and Fehr, 2011; Sacks, 1992; Schegloff, 2007). An *ask* is closely related to the notion of a request (Zemel, 2017). Perhaps most importantly, an ask elicits relevant responses from the recipient. *Framing* refers to linguistic and social resources used to persuade the recipient of an ask to comply and perform the requested social action. Put another way, an ask creates a social obligation to respond, while framing provides an adequate basis for compliance with the ask.
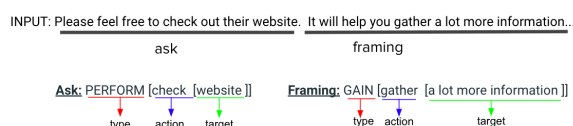

Figure 2: Example of ask and framing representations used as training for generation of Response Plans.

In Fig. 2, we show the ask/framing representational formalism that serves as the basis of our response plans. Here the *ask* is a request to PERFORM the action of *check out the website*. The perceived risk or reward (or *framing*) for this request is that, upon performing the action, one may GAIN something, i.e., *gather a lot more information*. We use two types of *asks*: GIVE (provide something or information) and PERFORM (perform an action), and two types of *framings*: GAIN (gain some benefit) and LOSE (lose benefit or resource). This preliminary ontology was motivated by conversa-
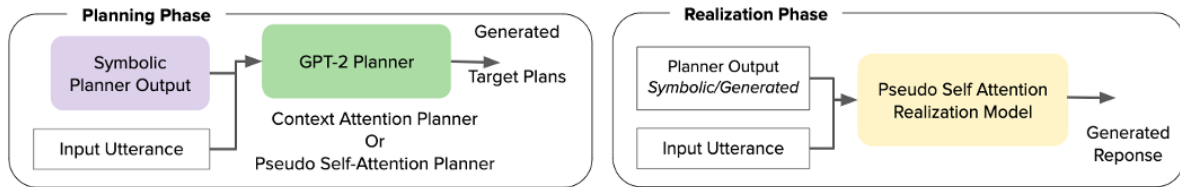
Figure 3: Architecture diagram of our system consisting of two phases: Planning and Realization. The Planning phase (Context and Pseudo Self Attention) encodes the input sequence and symbolic planner input to produce the response plans. The Realization phase uses the response plan and input utterance to generate the response

tion analysis literature (Sacks et al., 1978; Curl and Drew, 2008; Epperson and Zemel, 2008): by treating utterances as actions, we are able to establish what each utterance seeks to accomplish and how a sender motivates the recipient in terms of the benefits and costs of compliant responses.

3.2 Method

Our goal is to generate an informative response to the input utterance by first generating an appropriate **Response Plan**. We train two components separately (c.f. Fig. 3). In the ***Planning Phase***, we experiment with generating plans in three ways:

1. Symbolic Planner: Foremost, we need to extract plans automatically from utterances. To accomplish this goal, our symbolic planner adapts lexical representations previously used for language analysis (Dorr et al., 2020) to the problem of constructing **Response Plans**. We use lexical conceptual structures and basic language processing tools (Gardner et al., 2017; Manning et al., 2014) for parsing the input, identifying the main **action**, identifying the arguments (or **targets**), and applying semantic-role labeling. Fig. 2 presents ask/framing examples (type, action and target).

Once response plans are identified for all utterances in a given corpus using the symbolic planner, we need to address *automated generation* of such plans. Using the asks and framings as annotated data for a "silver" standard,[1] we train models to learn to generate "Response Plans" that are encoded with the same representation format used for asks/framings. We use the language modeling paradigm and use a large pre-trained model (GPT-2) (Radford et al., 2019) with the transformer architecture and the self-attention mechanism (Vaswani et al., 2017). We fine-tune this language model with the constraint of the input utterance and the plan for this input utterance, and train it to pro-

duce the plan for the response utterance. We adopt the fine-tuning approach specified by Ziegler *et al.* (2019) and train two specific models (CTX and PSA) described below.

2. Context Attention Planner (CTX): based on the encoder/decoder architecture. In this model, the decoder weights are initialized with the pre-trained weights of the language model. However, a new context attention layer is added in the decoder that concatenates the conditioning information to the pre-trained weight. The conditioning information, in our case, is the plan for the input utterance.

3. Pseudo Self Attention (PSA): Proposed by Ziegler *et al.* (2019), PSA injects conditioning information from the encoder directly into the pre-trained self attention (similar to the "zero-shot" model proposed by Radford *et al.* (2019)).

In the ***Realization Phase***, we generate responses by utilizing the response plan generated from the planning phase as well as the input utterance. We expect a more guided generation of responses that are constrained by the response plan. In this phase, we only experiment with the Pseudo Self attention (PSA) model, based on Ziegler *et al.* (2019), who demonstrate that PSA outperforms other approaches on text generation tasks. We use nucleus sampling to overcome some of the drawbacks of beam search (Holtzman et al., 2020).

3.3 Corpora

Our choice of corpora is driven by the presence of information elicitation and persuasive strategies in the utterances (i.e., asks and framings).

Accordingly, we experiment with the AntiScam (Li et al., 2019) and Persuasion for Social Good (Wang et al., 2019) corpora. **AntiScam** contains dialogues about a customer service scenario and is specifically crowdsourced to understand human elicitation strategies. **Persuasion for Social Good** corpus contains interactions between workers who are assigned the roles of persuader and persuadee,

---

[1]Dorr *et al.* (2020) report precision of 69.2% in detecting asks/framings.

	AntiScam	PSG
Number of Dialogues	220	1017
Avg. Conversation Length	12.45	10.43
Avg. Utterance Length	11.13	19.36
Number of GIVE	2192	11587
Number of PERFORM	1681	7335
Number of GAIN	70	399
Number of LOSE	73	588
Number of RESPOND	4376	8078

Table 1: Statistics of AntiScam and Persuasion for Social Good (PSG), with annotated asks and framings. Avg. conversation length - average number of turns in each conversation; Avg. utterance length - average length of a turn in a conversation

where the persuader attempts to convince the persuadee to donate to a charity.

All utterances in these corpora are first annotated through the Symbolic Planner (c.f. Section 3.2) to gauge suitability based on the presence of asks and framings. In Table 1, we provide descriptive statistics of the corpora; we find an adequate number of ask/framing types (GIVE, PERFORM, GAIN, LOSE). In cases where there are no asks/framings or the symbolic planner fails to detect them, we use the default action RESPOND.

3.4 Implementation

We implement the models using Open-NMT (Klein et al., 2017) and the PyTorch framework.[2] We use publicly available GPT-2 model (Radford et al., 2019) with 117M parameters, 12 layers and 12 heads in our implementations. The input utterances and the plans are tokenized using byte-pair encoding to reduce vocabulary size (Sennrich et al., 2015). Both phases are trained separately. In the Planning Phase, the *plan for the input* utterance along with the input utterance is used to generate the *response plan* for the response utterance; in the Realization Phase, the response plan and input utterance are input to the model to generate the response. In both planning and realization phase, separation tokens are added (e.g. <plan>), as is common practice for transformer inputs (Devlin et al., 2019; Wolf et al., 2019). We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0005 and $\beta_1 = 0.9$ and $\beta_2 = 0.98$. During decoding, we use nucleus sam-

---

[2] https://pytorch.org/

pling both in the planning and realization phase. All models are trained on two TitanV GPU and take roughly 15 hours each to train the planner and realization component. The trained models and the codebase are available at https://github.com/sashank06/planning_generation

4   Evaluation of Approach

The results reported in these subsections were obtained by combining both corpora and dividing randomly in a ratio of 80/10/10 for the training, testing, and validation set.

4.1   Planning Phase Evaluation

This evaluation focuses on investigating the efficacy of the two automated planners (Context Attention (CTX) and Pseudo-Self Attention (PSA)) in learning to generate response plans.

4.1.1   Automated Metrics

Are the automated planners able to faithfully learn how to generate the response utterance plans? To investigate, we compare the performance of the CTX and the PSA planner with the symbolic planner output (which is our silver standard reference) using common automated metrics Table 2: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) on the test set. We use the library by Sharma *et al.* (2017). We find that PSA was able to achieve higher word overlap metrics with respect to the silver standard. We conducted an in-depth analysis of the CTX and PSA planner output on the entire testing set. We found that the PSA model was more likely to produce ask actions that matched the ground truth, resulting in higher scores on the automated metrics.

4.1.2   Human Evaluation

Evaluation using automated metrics provides limited evidence for the ability to automatically generate plans; we do not know if these plans are actually useful in a realization task. The question then is: *How well-suited are the automatically learned plans for the task of generating responses?*

Study 1: We asked two experts in linguistics to independently rate 40 randomly sampled plans from the test set. For context, we provided the input utterance and its plan produced by the symbolic planner. Their task was to choose which of the learned response plans was better suited to the realization task (CTX, PSA, Both or Neither). They

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	ROUGE@L	METEOR
Context Attention (CTX)	0.1097	0.0714	0.0571	0.0506	0.5053	0.1677	0.3444
Pseudo-Self Attention (PSA)	**0.1342**	**0.0886**	**0.0672****	**0.0578****	**0.6506**	**0.2108**	**0.3447**

Table 2: Automated Metrics on performance of models in the Planning Phase. ** indicates $p < 0.01$

	CTX	PSA	Both	Neither
Q1	38.75%	26.25%	25%	10%
Q2	27.5%	20%	23.75%	28.75%
Q3	22.5%	17.5%	41.25%	18.75%
Q4	32.5%	31.25%	10%	26.25%

Table 3: Human Evaluation results on the performance of the planner component. **Q1:** Which model plan is better suited for generating a response?; **Q2:** Which model has the more appropriate ask/framing type?; **Q3:** Which model has the more appropriate ask/framing action with respect to the type?; **Q4:** Which model has the more informative ask/framing target?

also evaluated the plan constituents: (**type**, **action** and **target**). We randomized the presentation order of the planner outputs across questions to avoid ordering/learning effects (Medin and Bettger, 1994). We find an inter-rater agreement (Shrout and Fleiss, 1979) of 0.5 ($p < 0.001$) between the linguists.

Table 3 shows the results from Study 1. From **Q1**, we find that CTX planner is better suited to generate an appropriate response over the PSA planner. Similarly, through **Q2**, **Q3**, and **Q4**, we find that the CTX planner is better able to generate the appropriate ask/framing types, actions, and targets. We also find that the linguists rated Neither plan was suited to generate a response 10% of the time. Put differently; the automatically generated plans would work 90% of the time to generate an appropriate utterance in the realization phase. The learned plans have trouble associating an appropriate ask/framing type and target (28.75% and 26.75%) but perform better with the ask/framing action (18.75% Neither rating).

This evaluation compares the automatic planners against one another, but *how well do the planners compare to the silver standard (symbolic planner)?*

Study 2: We asked the same linguistic experts to independently determine which amongst two plans (symbolic vs. each automated planner) would be more appropriate to generate a response. This study design is consistent with prior studies in dialogue evaluation (Mei et al., 2017; Serban et al., 2016).

	CTX	PSA	Symbolic Planner	Both
Quality	30%	X	35%	35%
	X	35%	22%	43%

Table 4: Human evaluation results comparing CTX and PSA planner separately to the Symbolic Planner

Table 4 presents the results from Study 2.

We find that experts prefer the plans produced by the symbolic planner over the CTX output but not over the PSA planner output. Inter-annotator agreement (Shrout and Fleiss, 1979) between the experts for this study was 0.54. While Study 1 compared CTX and PSA planner outputs against one another, Study 2 compared CTX and PSA outputs against the silver standard. As we observe from the automated metrics (Table 2), PSA model plans are more faithful to the ground truth, e.g., higher BLEU 1-4 scores than CTX model plans. Since PSA planner outputs are more faithful to the ground truth, this may be why human judges rate them as preferable more often when compared against ground truth.

Planning Phase Evaluation Findings: To summarize this evaluation section, we find: PSA outperforms the CTX planner on automated metrics. This finding is consistent with the results from Ziegler *et al.* (2019). From Study 1, we find that both the planners are able to generate appropriate plans, with the appropriate ask/framing type, action, and target for the realization phase, a large proportion of the time. From Study 2, we find that when compared to the silver standard plans, PSA planner output is preferred over the CTX planner.

4.2 Realization Phase Evaluation

While the previous section focuses on evaluating the ability to generate plans automatically, we do not yet know *whether separating the generation process into planning and realization produces better responses than an end-to-end system?*

Thus, we compare four approaches towards realizing a response given an input utterance (through the Pseudo-Self Attention fine-tuned realization algorithm): (1) **No Planner** model which receives

Realizer Input	Dataset	BLEU	Diversity	Length	BERT-score
No Plan	AntiScam	0.0658	**0.0067**	7.168	0.841
	PSG	0.1149	**0.0049**	13.713	0.845
Symbolic Planner	AntiScam	**0.1814**	0.0062	6.245	**0.844**
	PSG	**0.1992**	0.0038	11.982	**0.848**
Context Attention Planner	AntiScam	0.0705	0.0064	7.298	0.84
	PSG	0.1027	0.0043	14.088	0.847
Pseudo Self Attention Planner	AntiScam	0.0692	0.0065	**7.553**	0.838
	PSG	0.1253	0.0045	**15.128**	0.847

Table 5: Automated metric results on the responses generated on the test set of both corpora.

input utterance but no plan as input; (2) **Symbolic Planner based Generation**: This model receives the plan from symbolic planner output; (3) **CTX Planner-Based Generation**: This model receives the CTX plan; (4) **PSA Planner-Based Generation**: This model receives the PSA plan.

4.2.1 Automated Metrics

Prior research has shown that most automated metrics have little to no correlation to human ratings on NLG tasks (Liu et al., 2016; Santhanam and Shaikh, 2019); however, they may provide some standard of reference to evaluate performance. We report the following metrics: (i) BLEU (Papineni et al., 2002) (ii) length of responses, with the understanding that models that are able to generate longer responses are better (iii) following, Mei *et al* (2017), we report the diversity metric (Li et al., 2016a). Diversity is calculated as the number of distinct unigrams in the generation scaled by the total number of generated tokens (Mei et al., 2017; Li et al., 2016b). (iv) BERT-Score (Zhang* et al., 2020) metric, an embedding-based score which has shown greater correlation to human ratings.

Table 5 reports on the automated evaluation against the ground truth utterances. We find that on both corpora and across all metrics except Diversity, incorporating plans as an additional input to the realization phase helps achieve a higher score than having No Planner. From Table 5, we find that the realizer without any plans is able to achieve higher diversity, but the difference is not statistically significant.

4.2.2 Human Evaluation

Since automated metrics are not the most informative indicators of quality of generated responses, thorough human evaluation is necessary. We *in-*

vestigate if humans prefer the responses generated by the planner-based models over those generated without the plan (No Planner). We conducted two human evaluation studies by recruiting workers from Amazon Mechanical Turk service with strict quality control criteria: workers should have at least 90% HIT approval rate and at least 1000 approved HITs. In each survey, workers are asked to evaluate responses on these metrics, following Novikova *et al.* (2018): (i) *Appropriateness:* determines whether response aligns with the topic of the conversation and the input utterance. (ii) *Quality:* determines the overall quality in terms of grammatical correctness, fluency, and adequacy (iii) *Usefulness:* determines if the response is highly informative to generate a response.

Realizer Input	Appropri-ateness	Quality	Useful-ness
No Plan	2.54	2.61	2.58
Symbolic Planner	2.51	2.5	2.53
CTX Planner	**2.34**	**2.38**	**2.38**
PSA Planner	2.59	2.5	2.51

Table 6: Average ranking of realized output from four different planners, lower score is better

Study 1: We tasked 30 crowd-sourced workers to rank order the four model responses from best to worst. We randomly sampled 60 examples from the test set with an even 50% split (30 examples each) between the Persuasion for Social Good and Anti-Scam corpora. We chose the best to worst ranking mechanism since it has shown greater consistency and agreement amongst workers on tasks related to dialogue evaluation over other evaluation designs (e.g. Likert scales) (Santhanam et al., 2020; Kiritchenko and Mohammad, 2017). The presentation

METRIC	Ground Truth	Symbolic Planner	TIE	Ground Truth	CTX Planner	TIE	Ground Truth	PSA Planner	TIE
COMPARISON OF PERFORMANCE - OVERALL									
Appropriateness	53	26	20	41	39	20	42	38	20
Quality	50	32	19	45	40	15	43	44	13
Usefulness	50	32	18	44	42	14	43	43	14
COMPARISON OF PERFORMANCE - PERSUASION FOR SOCIAL GOOD CORPUS									
Appropriateness	58	24	18	47	33	20	40	37	23
Quality	55	30	15	53	31	16	39	46	15
Usefulness	55	33	12	50	35	15	40	44	15
COMPARISON OF PERFORMANCE - ANTI-SCAM CORPUS									
Appropriateness	48	29	23	35	45	20	44	38	18
Quality	44	34	22	37	48	15	47	42	11
Usefulness	46	31	23	37	48	15	45	41	14

Figure 4: Comparison of ground truth reference with realized output from each model that receives learned plans as input: Symbolic, CTX or PSA. Higher values (shown as %)/darker color represent better performance.

order of model outputs for each question was again randomized to avoid learning effects (Medin and Bettger, 1994). Table 6 demonstrates the average rank position (1=Best, 4=Worst) obtained by each model. We find using the plans generated by the CTX planner helps generate better responses. On the metrics of quality and usefulness, we find that incorporating planning as additional input performs better than no plan (i.e. end-to-end system).

Study 2: In this study, we evaluate *how well the generated responses compare to the ground truth.* The ground truth references are those produced by humans in the PSG and Anti-Scam corpora. We recruited 11 MTurk workers with the same crowdsourcing quality controls as Study 1. For the same randomly sampled 60 examples from Study 1, workers were asked if they prefer the ground-truth response, the response generated from the three planners, or both, on the three chosen metrics. This study design is also consistent with prior work (Mei et al., 2017). Workers were blinded to the source of the response (ground truth or generated) and were presented the responses in a randomized order across all questions to avoid ordering effects.

Fig. 4 shows the results (higher value/darker color is better): we find that responses generated from the symbolic planner as input do not perform well when compared to the ground truth. In other words, the proportion of time that the ground truth response is preferred over that generated by the symbolic planner is significant (e.g. 53% vs. 26% on the Appropriateness metric overall).

We find that on all three metrics, the responses generated using CTX and PSA plans help generate responses that are comparable to the responses produced by humans (ground truth). We also find that the PSA planner-based responses perform better overall and on the Persuasion for Social Good

Figure 5: Sample outputs from realization phase with all variations of planner input, as well as the ground truth response from the corpus

corpus. Surprisingly, the CTX planner based responses performs better than Ground Truth utterances for the Anti-Scam corpus (45%, 48% and 48% of the time preferable vs. ground truth response 35%, 37% and 37% on the three metrics, Appropriateness, Quality, and Usefulness, resp.). We explain this unexpected finding in the next subsection (Section 4.3).

Realization Phase Evaluation Findings: To summarize this evaluation subsection, we find that the Symbolic Planner-realized output outperforms the CTX, PSA, and No Planner output on the automated metrics of BLEU and BERT-score. Importantly, the CTX planner-realized output has a higher rank in terms of overall preference in human evaluation than the other models (c.f. Table 6). We also find that human-generated utterances (ground truth) are preferred overall (c.f Fig. 4) than the model outputs. We found inter-rater consistency and agreement scores to be >0.6 on average across the metrics (full tables are reported in the Appendix).

Issues Found	Input Utterance and Context	Generated Plan for Response/ Generated Response
	Planning Phase	
Non-Informative Ask/Framing Target	The money goes directly to the organization in order to help. The money will have a large impact. Would you like to have the site info?	PERFORM [see [that]]
Wrong Ask/Framing & type and action	no, that has never been my address.	GIVE [give [why got]]
	Realization Phase	
Ignored Plan	you certainly can, I am also able to take this payment information as well.	Plan: PERFORM [find [your billing date and names]] Response: Okay, thanks!
Grammatical Inconsistency	Person A: The payment has not been processed. Person B: Oh why not?	Plan: PERFORM[help[with that]] Response: They help with that.

Table 7: Summary table of issues found through a qualitative analysis of the planning and realization phase outputs

4.3 Qualitative Analysis

We conduct a qualitative evaluation of the outputs and present several cherry- and lemon-picked examples here. Additional examples of success and failure cases are uploaded in the Appendix. In the sample conversation shown in Figure 5, we find that realized outputs using CTX and PSA plans are more consistent with the context of conversation than the symbolic planner approach. Additionally, the No Planner output (an end-to-end system which does not get a plan as an additional input) produces an utterance that may not necessarily continue the conversation further.

This example is also illustrative of the finding in Study 2 of the Planning Phase evaluation, where the crowdsourced workers rated the automated planner-based outputs better than the symbolic planner-based outputs (c.f. Fig. 7). This might seem contradictory, as the CTX and PSA planners are trained on the silver standard data from the symbolic planner. We contend that this is due to the ability of automated planners (CTX and PSA) to generalize, an ability lacking in the symbolic planner. In such cases, as shown in Fig. 5, the symbolic planner defaults to the RESPOND message plan, and this lead to generated output: *That is not an exact word*, which is generic and off-topic. The symbolic planner could be improved to cover more cases; however, the effort would not be scalable.

While we find promising results for the automatically-generated planners in Sections 4.1 and 4.2, areas of improvement do exist (Table 7):

Non-Informative Ask/Framing Targets: We find several examples where the ask/framing tar-gets are non-informative words (e.g. *this, that*). Non-informative targets can cause the downstream realization process to generate an utterance that is, in turn, also non-informative. One example of such cases is shown in Row 1 of Table 7.

Wrong Type and Action: Another planning phase issue category is that the constituents of plan representation (e.g., the ask/framing type and action) can be incorrect. As illustrated by the example in Table 7, an ask target of *why got* is incorrect. Typically, we would expect to find a noun or a noun phrase as the ask/framing action (e.g., *your billing date and names* as shown in the plan in Row 3).

Ignored Plan: In the Realization phase, a typical issue is that the realizer may ignore the generated plan. As can be seen in Row 3 of Table 7, the plan should constrain the response, and thus should contain phrases such as *finding your billing date and names*. However, the generated response is instead a generic phrase *Okay, thanks!*.

Grammatical inconsistencies: We also note that there were cases where the grammar, e.g. pronoun usage, is inconsistent. For the example shown in Row 4 of Table 7, we see that the generated response is *They help with that.* whereas the conversation is between two persons; a generated response of *I can help with that* would be more consistent with the context of the conversation.

5 Conclusion and Future Work

We address the task of natural language generation in open-ended dialogue systems. We test our hypothesis that decoupling the generation process into planning and realization can achieve better

performance than an end-to-end approach.

In the planning phase, we explore three methods to generate response plans, including a Symbolic Planner and two learned planners, the Context Attention and Pseudo Self Attention models. Through linguist expert evaluation, we are able to determine the efficacy of the response plans towards realization. *In the realization phase*, we use the Pseudo Self Attention model to make use of the learned response plans to generate responses.

Our key finding through two separate human crowdsourced studies is that decoupling realization, and planning phases outperforms an end-to-end No Planner system across three metrics (Appropriateness, Quality, and Usefulness).

In this work, we have taken an initial step towards the goal of replicating human language generation processes. Thorough and rigorous evaluations are required to fully support our claims, e.g., by including additional metrics and more diverse corpora. In this work, we limit the types to GIVE, GAIN, LOSE, and PERFORM. However, we do not restrict the ask action and target at all. Also, since our symbolic planner can be used to obtain silver standard training data, straightforward changes like adding additional lexicons would enable us to generalize to other corpora as well as include additional ask types in our pipeline. Another natural extension would be to explore training the planning and realization phases together in a hierarchical process (Fan et al., 2018). This would, in principle, further validate the efficacy of our approach.

Acknowledgments

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Kathryn Bock. 1996. Language production: Methods and methodologies. *Psychonomic Bulletin & Review*, 3(4):395–421.

Kathryn Bock, Gary S Dell, Franklin Chang, and Kristine H Onishi. 2007. Persistent structural priming from language comprehension to language production. *Cognition*, 104(3):437–458.

Philip R Cohen and C Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive science*, 3(3):177–212.

Traci S Curl and Paul Drew. 2008. Contingency and action: A comparison of two forms of requesting. *Research on language and social interaction*, 41(2):129–153.

Gary S Dell. 1985. Positive feedback in hierarchical connectionist models: Applications to language production 1. *Cognitive Science*, 9(1):3–23.

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.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Bonnie J Dorr, Archna Bhatia, Adam Dalton, Brodie Mather, Bryanna Hebenstreit, Sashank Santhanam, Zhuo Cheng, Samira Shaikh, Alan Zemel, and Tomek Strzalkowski. 2020. Detecting asks in se attacks: Impact of linguistic and structural knowledge. *arXiv preprint arXiv:2002.10931*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. *arXiv preprint arXiv:1903.09722*.

Terrence W Epperson and Alan Zemel. 2008. Reports, requests, and recipient design: The management of patron queries in online reference chats. *Journal of the American Society for Information Science and Technology*, 59(14):2268–2283.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Konstantina Garoufi. 2014. Planning-based models of natural language generation. *Language and Linguistics Compass*, 8(1):1–10.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Shelia M Kennison. 2018. *Psychology of Language: Theory and Applications*. Macmillan International Higher Education.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2019. End-to-end trainable non-collaborative dialog system. *arXiv preprint arXiv:1911.10742*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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 *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Douglas L Medin and Jeffrey G Bettger. 1994. Presentation order and recognition of categorically related examples. *Psychonomic bulletin & review*, 1(2):250–254.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2017. Coherent dialogue with attention-based language models. In *AAAI*, pages 3252–3258.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. 2017. "how may i help you?" modeling twitter customer serviceconversations using fine-grained dialogue acts. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 343–355.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4526–4536, Hong Kong, China. Association for Computational Linguistics.

Anita Pomerantz and B. J. Fehr. 2011. Conversation Analysis: An approach to the analysis of Cocial interaction. In van Dijk, Teun., editor, *Discourse Studes: A Multidisciplinary Approach*, pages 165–190. Sage.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6908–6915.

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

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5370–5381.

Ehud Reiter. 1994. Has a consensus nl generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 163–170. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Harvey Sacks. 1992. *Lectures on Conversation, Volumes 1 & 2*. Blackwell.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590, Hong Kong, China. Association for Computational Linguistics.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 673–683. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5635–5649. Association for Computational Linguistics.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Wei Wu, Can Xu, Yu Wu, and Zhoujun Li. 2018. Towards interpretable chit-chat: Open domain dialogue generation with dialogue acts.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Idris Yusupov and Yurii Kuratov. 2018. NIPS conversational intelligence challenge 2017 winner system: Skill-based conversational agent with supervised dialog manager. In *Proc. 27th International Conference on Computational Linguistics*, pages 3681–3692, Santa Fe, New Mexico, USA. Assoc. for Computational Linguistics.

Alan Zemel. 2017. Texts as actions: Requests in online chats between reference librarians and library patrons. *Journal of the Assoc. for Information Science and Technology*, 67(7):1687–1697.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Li Zhou, Kevin Small, Oleg Rokhlenko, and Charles Elkan. 2017. End-to-end offline goal-oriented dialog policy learning via policy gradient. *arXiv preprint arXiv:1712.02838*.

Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.

A Supplementary Materials

A.1 Planner Output Analysis

Table 8 shows the performance of planners on the test set. We count the number of ask/framing types as well as the number of default response plan produced by each planner: GIVE, PERFORM, GAIN, LOSE and RESPOND from the testing set.

	Symbolic Planner	CTX Planner	PSA Planner
Num of GIVES	1248	1187	1146
Num of PERFORM	815	1041	1129
Num of GAIN	44	35	29
Num of LOSE	66	35	49
Num of RESPOND	969	842	789

Table 8: Distribution of different types of asks/framings in the test set of the planning component. Note: We found two asks produced by the CTX planner that ignored our ontology, which are excluded from our counts in this table

A.2 Inter-rater Consistency for Realization Phase

We present the inter-rater consistency and agreement scores for the crowd-sourced worker studies we conducted during Realization Phase evaluation. The results presented were calculated using the R *irr* package.[3]

	Appropriateness	Quality	Usefulness
Consistency	0.42	0.65	0.67
Agreement	0.42	0.65	0.67

Table 9: ICC-Consistency and Agreement Scores on the Appropriateness, Quality and Usefulness metrics obtained from Study 1 in the Realization Phase

	Appropriateness	Quality	Usefulness
Consistency	0.65	0.60	0.58
Agreement	0.62	0.59	0.55

Table 10: ICC-Consistency and Agreement Scores on the Appropriateness, Quality and Usefulness metrics obtained from Study 2 in the Realization Phase

---

A.3 Additional Output Examples

In this section, we give additional examples of conversations from our test set. Realization output based on each planner configuration of the system is included. In the Tables 11, we provide additional examples of the issues we found through manual inspection of the outputs.

Figure 6: Example conversation between two speakers A & B from test set. We can see that in the case of the realizer output from Symbolic Planner and PSA Planner, the responses also includes an ask (e.g. *what is that for?* and *Are you involved with them?* which may serve to carry the conversation further, as compared to other responses.

Figure 7: Another conversation between two speakers A & B from our test set. The ground truth response in this case was lengthier than the typical response (consisting of 73 words) and has been shortened here for ease of presentation.

Conversation context and input

Person A: wow, yeah I know that is horrible. Please do not judge me for saying that i can only donate .25 cents

Person B: I would never judge you for only donating .25 cents! That's .25 more cents than they had before. Just think that could be your child... Wouldn't you want any funding you could receive so your child could benefit!

Ground Truth Response
Person A: yes I would. And you are very kind!! the way you described their scenery really is eye opening. poor babies, your blessed for helping

No Planner
Person A: ok, i will donate 30 cents

Symbolic Planner
Person A: That is a very generous sentiment, but i really appreciate you on the other side. That's, how are you helping?

CTX Planner
Person A: I did not know that. If I donate if it is ethical organization, I will raise awareness about it.

PSA Planner
Person A: Ok, it's not like this amount just for us, but to the family and the community, it makes me feel like this is a particular matter of opportunity. I dont button pop up to donate, but to myself we have to do something. Okay.
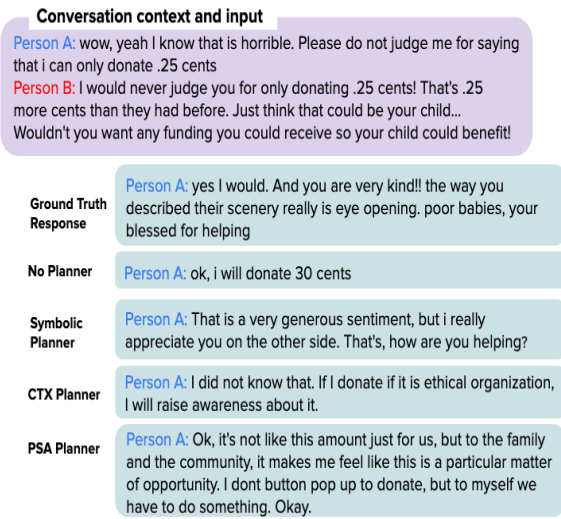
Figure 8: Third conversation example from the test set. In this example, we noticed that the PSA planner based model generates an incoherent response.
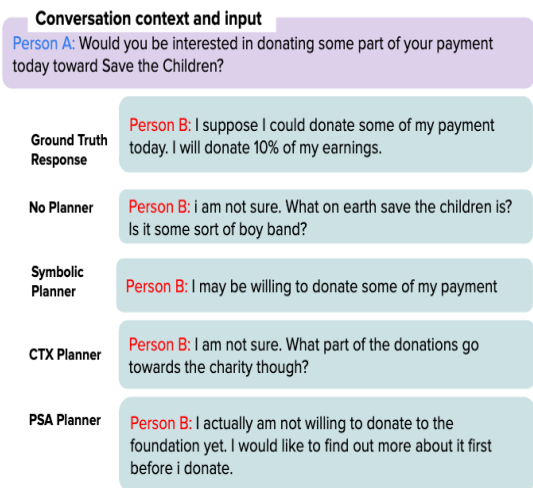
Conversation context and input

Person A: Would you be interested in donating some part of your payment today toward Save the Children?

Ground Truth Response
Person B: I suppose I could donate some of my payment today. I will donate 10% of my earnings.

No Planner
Person B: i am not sure. What on earth save the children is? Is it some sort of boy band?

Symbolic Planner
Person B: I may be willing to donate some of my payment

CTX Planner
Person B: I am not sure. What part of the donations go towards the charity though?

PSA Planner
Person B: I actually am not willing to donate to the foundation yet. I would like to find out more about it first before i donate.

Figure 9: An additional conversation example with output from realizers with all four planner configurations.

Issues Found	Input Utterance and Context	Generated Plan for Response/ Generated Response
	Planner Phase	
Non-Informative Ask/Framing Target	I donate pretty regularly with my kids in school, they have different charities they help out each month.	PERFORM [look [i]]
Non-Informative Ask/Framing Target	If you get time, go to Save the Children's website and check out more information. Maybe someone you know would like to make a donation or match your donation to help children in other countries affected by war and poverty.	PERFORM [support [them]]
Non-Informative Ask/Framing Target	Thats great to hear, empathy and sympathy goes a long way and is lacking in society so its great to hear another parent trying to help out as much as they can.	GIVE [donate [to]]
Wrong Ask/Framing & type and action	Have you felt inclined to donate to any charity before? If so, what sort of charity?	GIVE [give [part]]
Wrong Ask/Framing & type and action	I don't understand. They make money through online searches?	GIVE [give [online]]
	Realization Phase	
Ignored Plan	Would you be interested in donating some of your payment today toward Save the Children?	Plan: GIVE[donate[$ 1]] Response: I am not sure. What part of the donations go towards the charity though?
Ignored Plan	Person A: Do you have a personal connection to the cause? Person B: Yes. Family members and friends have gone through this.	Plan: LOSE[lost [my grandmother]] Response: I have a friend that lost his parents at the Mexico-US border. It broke my heart
Grammatical Inconsistency	Person A: i want to assure you we are on an encrypted line.	Plan: PERFORM[verify[that]] Response: i can verify that, please verify that.
Grammatical Inconsistency	Person A: could you verify your credit card expiration date Person B: what does it say on your end?	Plan: PERFORM[provide [that i need to verify it from you]] Response: i prefer to provide that info, so please provide that info.

Table 11: Summary table of issues found through a qualitative analysis of the planning and realization phase outputs