INTERVIEW: Large-scale Modeling of Media Dialog with Discourse Patterns and Knowledge Grounding

Bodhisattwa Prasad Majumder* **Shuyang Li***
Jianmo Ni† **Julian McAuley**
Computer Science and Engineering
University of California, San Diego
{bmajumde, shl008, jin018, jmcauley}@ucsd.edu

Abstract

In this work, we perform the first large-scale analysis of discourse in media dialog and its impact on generative modeling of dialog turns, with a focus on interrogative patterns and use of external knowledge. Discourse analysis can help us understand modes of persuasion, entertainment, and information elicitation in such settings, but has been limited to manual review of small corpora. We introduce INTERVIEW—a large-scale (105K conversations) media dialog dataset collected from news interview transcripts—which allows us to investigate such patterns at scale. We present a dialog model that leverages external knowledge as well as dialog acts via auxiliary losses and demonstrate that our model quantitatively and qualitatively outperforms strong discourse-agnostic baselines for dialog modeling—generating more specific and topical responses in interview-style conversations.

1 Introduction

Much of the news, information, and punditry the general public listens to and reads consists of *media dialog*—a category of open-domain conversations between an interviewer and interviewee centered on world events and situational context. A system for modeling media dialog from the perspective of one of these roles can help us better understand how media persuades and informs the public (Southwell et al., 2018). Thus, while recent work in dialog modeling has focused on goal-oriented (Bordes et al., 2017), spontaneous (Shao et al., 2017), or synthetic open-domain chit-chat (Li et al., 2017; Dinan et al., 2019; Gopalakrishnan et al., 2019), we aim to analyze discourse patterns in media dialog and their impact on dialog modeling.

Media dialog differs linguistically and in purpose from unstructured, spontaneous conversation such as open-domain chitchat, and both the topical content and interlocutor intent are heavily influenced by the social, cultural, and temporal setting (Weizman, 2008). The study of media dialog has traditionally focused on individual and manual review of small-scale (<200K word) news corpora (Bednarek, 2006; van Dijk, 2011), and we see an opportunity to scale some forms of discourse analysis to tens of thousands of such documents. In this work, we perform the first large-scale automatic analysis of structural components (response-type patterns) and question type categorization on media dialog, specifically for English news interviews. We show that predicting discourse features can improve generative dialog modeling performance, demonstrating the degree to which discourse structure impacts an interviewer's choice of response type and content. News interviews are also heavily situation-grounded and contextualized by past events and world knowledge. We explore methods to associate each conversation with a selection of world facts, and show that by modeling interviewers as *knowledge-grounded* speakers mediating a conversation we are able to generate relevant and specific utterances fitting their role.

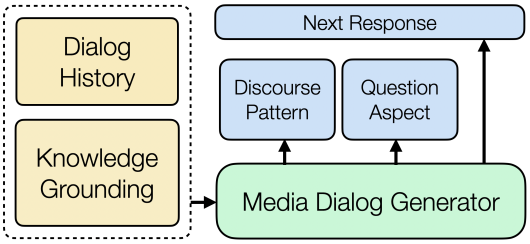


Figure 1: Our dialog model incorporates grounding documents alongside dialog history. We also leverage the dialog patterns and interrogative positioning by the host via auxiliary losses.

Our main contributions in this work are:

1. We collect a dataset of 105K media dialogs

* denotes equal contribution
† Now at Google

(23K two-party dialogs)[1] encompassing two decades of National Public Radio (NPR) radio programs, on which we conduct extensive experiments;

2. We present a probabilistic framework to link a dialog with facts from a large corpus of grounding documents and show that it improves downstream dialog modeling performance compared to a strong TF-IDF baseline;

3. We introduce two auxiliary losses to guide utterance generation in a media dialog setting: look-ahead dialog structure prediction and question-attribute prediction[2]. We show that these losses significantly improve generation quality via automatic and human metrics.

2 Related Work

Media dialog—specifically, the news interview—has seen study primarily in the field of speech transcription, diarization, and speaker role modeling (Chen et al.; Laurent et al., 2014). These works have typically focused on techniques to annotate broadcast audio transcripts (Hutchinson et al., 2010) in order to cluster different news stories from a continuous broadcast stream (Huang et al., 1999). While Barzilay et al. (2000) and Liu (2006) note that transition points between speaker roles (e.g. anchor and guest) can determine the high-level topical flow of a news conversation, we investigate the impact of discourse patterns on the semantics of specific utterances.

Such research is currently limited by a lack of accessible corpora for the study of media dialog at scale. The Defense Advanced Research Projects Agency has undertaken efforts to collect and transcribe broadcast conversations (Strassel, 2004; Cohen, 2007). However, it proves difficult to adopt these datasets as widely available benchmarks on dialog modeling tasks, as they come with a substantial cost ($100-$1000 per annum per dataset). More recent efforts to amass such data have either focused on collecting large volumes of conversation fragments with noisy transcripts (Beeferman et al., 2019) or human transcripts for a smaller set of long-form open-domain radio programs (Mao et al., 2020). We contribute an open-access large-scale corpus of broadcast media dialog annotated with response types, demonstrating that these are useful for modeling interviewer utterances.

We explore the application of discourse analysis (Fairclough and Wodak, 1997) on this large media dialog corpus in order to discover, confirm, and leverage *discourse patterns* regarding interrogative forms, speaker agency, and references to external knowledge. As noted by Weizman (2008) in their deep study of Israeli news television, structure in media dialog (in contrast to spontaneous natural conversation) is uniquely determined by its speaker role dynamics. Wang et al. (2011) investigate the detection of one such dynamic: agreement/disagreement between speakers. Ma et al. (2019) classify discourse relations (e.g. comparative, temporal) between two turns of dialog, but do not study discourse structure. In this work we extend our analysis to other properties of interviewer utterances (e.g. subjectivity, polarity, dialog act patterns) (Heritage, 1985) in the context of generative dialog modeling. Structured approaches for dialog modeling employ a simple concatenation of dialog history in a transformer-based architecture (Zhang et al., 2019). We draw inspiration from Luan et al. (2017) who demonstrate the usefulness of a multi-task framework for speaker-conditioned dialog modeling. Guu et al. (2020) propose a framework for jointly learning document retrieval and language modeling, and we propose a similar model to learn task-specific annotation of grounding documents.

3 INTERVIEW : A Media Dialog Corpus

We collect a new dataset of 105K multi-party interview transcripts for 7 programs on National Public Radio (NPR)[3] over 20 years (1999–2019). These transcripts contain in total 3M turns comprising 7.5M sentences (127M words) from 184K speakers, of which 287 are interviewers. To investigate host-mediated media dialog, we curate a subset, **INTERVIEW 2P**, with **two roles**: an *interviewer* and a *guest*, comprising 23K two-party conversations encompassing 455K turns, with 1.24M sentences and 21.7M words. In these two-party conversations, each speaker takes an average of nine turns per dialog. Guests tend to speak longer on their turns, with 1.6x as many sentences spoken and 2x as many words per turn. Meanwhile, hosts ask five times as many questions as guests, with 40% of their dialog turns containing questions. When ask-

[1] https://www.kaggle.com/shuyangli94/interview-npr-media-dialog-transcripts

[2] Code: https://github.com/MEDIA-DIALOG/interview-media-analysis

[3] https://www.npr.org/

Host (Question): Steve Bannon is quoted as saying [...] the president has lost it. Now, are you supporting a president who is incapable of being entrusted with [...] nuclear weapons?

Guest (Answer): Well - one thing I haven't heard yet is Steve Bannon interviewed […] so look, I think the president of the United States has shown he's very, very capable […]

Host (Question): Should he be taunting a dictator with nuclear weapons about the size of his nuclear button?
[Question types: Polar, Subjective, Combative]

Guest (Answer): Well, I - you know […] the president has a record on Twitter […] I think he makes points […] he's doing a great job from where I sit for the country.

Host (Question): Quickly, he says he's a genius. Do you agree?

Figure 2: Example conversation from INTERVIEW with annotated discourse analysis. Text highlighted in blue indicates the question of interest, uttered by the host. The dialog triplet is marked in red.

Dataset	Structured	# Dialogs	# Turns	# Words
RadioTalk (2019)	✗	5.98 M*	116 M	2.9 B
TAL (2020)	✓	663	163,808	7.4 M
INTERVIEW 2P	✓	23,714	454,739	21.7 M
INTERVIEW	✓	105,848	3,199,856	126.7 M

Table 1: Comparative media dialog dataset statistics. *RadioTalk does not contain full conversations

ing questions, hosts and guests use interrogative forms (See et al., 2019) at the same rate (65%).

3.1 Comparison with Other Datasets

Open-domain dialog datasets have traditionally focused on either spontaneous (e.g. telephone calls) or goal-oriented conversation, and there is a paucity of English-language *media dialog* datasets—that is, dialog corpora comprising semi-structured conversations for the purpose of information elicitation and presentation. The closest such datasets are This American Life (Mao et al., 2020), a dataset of several hundred long-form expository podcast episodes, and RadioTalk (Beeferman et al., 2019), which comprises over one million ten-minute snippets of talk radio transcripts. While these corpora are derived from broadcast media, episodes of the former contain a broad range of expository speakers who are not professional journalists, while the latter dataset is constructed via an automated transcription system with a 13%+ word error rate and does not contain full conversations (segments from radio conversations are transcribed). We compare INTERVIEW statistics to other English media dialog datasets in Table 7.

Traditional media dialogs (e.g. news interviews) comprise a significant body of media consumed by the general public and we believe there is value in the large-scale study of such media. Efforts to collect and transcribe broadcast news span the world, from the French EPAC corpus (Estève et al., 2010) to Arabic and Chinese news manually transcribed via the GALE program (Cohen, 2007). To our knowledge, no attempt has yet been made to analyze the discourse patterns or trends in such data—these datasets have primarily been used to support the development of automatic speech recognition, transcription, and machine translation systems. Early efforts to collect English-language broadcast conversation transcripts (Placeway et al., 1997) similarly aimed to build smaller, high-quality parallel corpora for speech transcription. The large-scale study of discourse in media dialog is not supported in such corpora, and the INTERVIEW corpus enables such analysis at scale for English-language media.

4 INTERVIEW Discourse Analysis

We tackle three aspects of discourse analysis that can be scaled to INTERVIEW: 1) Dialog patterns that emerge through new interviews; 2) Large scale annotation of interviewer question types (dialog acts); and 3) Obtaining grounding documents that provide situational context for a news interview. We study these discourse features in context of English broadcast news interviews.

4.1 Dialog Patterns

The news interview setting revolves around sets of questions and answers—naively, one may assume the interviewer to be the sole questioner. However, media dialog has steadily deviated from this rigid structure, tending toward the broadly conversational (Fairclough, 1988). Each participant may be at turns jovial, inquisitive, and critical, and this

is reflected in question-answer patterning. Heritage (1985) frames the analysis of media discourse in terms of the *third-turn receipt*, where 1) they ask a question; 2) the interviewee responds; and 3) the interviewer chooses how to proceed. We are motivated by this, as well as studies of *question-response-confirmation* patterns in spontaneous dialog (Van Hekken and Roelofsen, 1982). We focus on discourse patterns in **response type triplets** beginning with an interviewer (host) question.

We define a triplet as $\{r_1, r_2, r_3\}$ where the response type at utterance i is a question or an answer: $r_i \in \{\text{Q}, \text{A}\}$. By imposing a binary label on each utterance, we are able to efficiently mine all occurrences of each of eight possible host-guest-host patterns across our 23K dialogs. We find that a structured interrogative Q-A-Q pattern comprises 27% of all cases, while 20% of the time the host poses a non-interrogative third response (Q-A-A). Guests respond to questions with questions of their own only 7% of the time, supporting the theory that interviewers serve as the primary *mediators* in such conversations (Weizman, 2008). Manual inspection evinces recurring action patterns corresponding to interviewer stance-taking and agendas ranging from cooperative to confrontational. For example, the conversation segment in Figure 2 is comprised entirely of Q-A-Q patterns, with the host prompting (Heritage, 1985) the guest, re-contextualizing and refocusing the guest's stance for the benefit of the audience. To leverage the inter-dependence of action choice (question or answer) and stance-taking (implicitly or explicitly via utterance content) (Haddington, 2004), we propose to predict the subsequent response type triplet while modeling an interviewer utterance. We thus explore how utterance phrasing and structure may depend on projected or desired conversation directions.

4.2 Question Types as Dialog Acts

In their role as a mediator, interviewers can shape the narrative by posing different *types* of questions to guests. Weizman (2008) posits that this choice of question type is influenced by dialog context and conversation flow. We examine ways to structurally bias our model to take advantage of conversational context in order to ask appropriate interviewer questions. Based on common interviewing guides[4] and linguistic analysis of open-ended questions in a conversational setting (Karttunen, 1977), we define three interrogative aspects (attributes): 1) **Polarity**: determining if the question is yes/no (polar) or open-ended; 2) **Subjectivity**: determining if it demands a factual answer or invites a subjective opinion; and 3) **Combativeness**: whether the question is confrontational or clarifying. Our mode of categorization resembles that of Gnisci and Bonaiuto (2003), who add additional categories that are more relevant to the study of equivocation in confrontational interviews. While previous works have primarily used question polarity and interrogative forms to improve diversity in spontaneous dialog generation (Zhao et al., 2017), we explore how a news interviewer constructs question contents given desired interrogative aspects.

History	Model	Polarity	Combativeness	Subjectivity
No	MLP	55.61	48.91	50.87
	CNN	68.20	57.19	53.91
	LSTM	66.87	49.70	51.96
	BERT	75.31	58.10	66.92
Yes	MLP	68.71	60.81	61.21
	CNN	74.71	65.87	67.98
	LSTM	70.49	60.54	63.09
	BERT	**80.20**	**70.14**	**76.92**

Table 2: F1 Performance of question-type classifier models on the test set.

We hired two expert annotators to assess a question based on these three aspects. We provided interviewer questions alongside corresponding dialog histories, and annotators marked the binary presence/absence of each aspect for each question. The first host question from Figure 2 would be marked as polar, subjective, and combative, as it asks the guest whether (polar) they endorse (subjective) an intentionally ridiculous statement (combative). We collected 1,000 questions in this manner, each labeled by both annotators. The inter-annotator agreement (Cohen's kappa (Cohen, 1960)) for each of the binary labeling tasks—polar vs. open-ended, subjective vs. objective, combative vs. clarifying—was 0.8 for polarity, 0.72 for subjectivity and 0.7 for combativeness. We observed questions in this sample to be 60.2% polar, 38.7% subjective, and 29.5% combative.

Automatic Classification We label the remainder of INTERVIEW by training a multi-label classifier, fine-tuning BERT (Devlin et al., 2019) to predict the presence of each attribute in our human-annotated set of questions. We concatenate dialog

[4] http://prndg.org/host-interviewing-tips

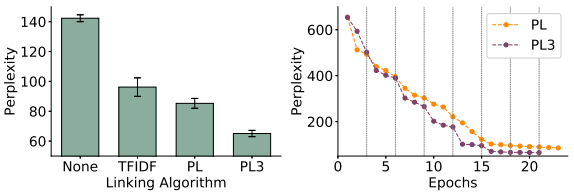
history and the interviewer question separated by a [SEP] token and prepend a [CLS] token. We calculate binary cross entropy loss over a linear projection of the final hidden state of the [CLS] token. BERT achieves 80.20, 70.14, and 76.92 F1 scores for polarity, combativeness and subjectivity respectively on the test set in four epochs.

We consider multiple baselines: 1) an MLP model using Bag-of-Words input features; 2) a CNN (Fukushima, 1988) with 2 convolution layers; and 3) a Bi-LSTM (Graves et al., 2005) network with max-pooling of final hidden layers. We initialize all embeddings with BERT embedding vectors. As shown in Table 2, BERT achieves the highest F1-score. Including dialog history improves classification performance, confirming that the type of question asked depends on conversational context. This suggests that we may also be able to better predict question content through jointly leveraging the dialog history and question type. Both human annotators and our model find predicting polarity the easiest, and combativeness the most difficult.

4.3 Knowledge Grounding

Media dialog is frequently characterized by references to world knowledge, current events, and factual information. This can be learned to some extent in large language models pre-trained on diverse text corpora (Petroni et al., 2019), and such models can act as knowledge stores (Chen et al., 2019). However, for tasks involving complex reasoning and induction it remains beneficial to provide models with externally linked knowledge (Mitra et al., 2019; Fan et al., 2019). Specifically for dialog modeling, the Wizard of Wikipedia (Dinan et al., 2019) and Topical Chat (Gopalakrishnan et al., 2019) corpora consist of grounding documents linked with open-domain chit-chat. As such, we explore methods to link *grounding knowledge documents* for each conversation in INTERVIEW, drawn from NPR news articles from the past two decades. We aim to link documents that can best inform conversation content and structure as measured by downstream dialog modeling performance.

TF-IDF Linking We assess a strong retrieval baseline for grounding document linking, using TF-IDF (Salton and Buckley, 1988) to find relevant documents for each conversation. To support large-scale TF-IDF similarity computation, we use the Lucene-based ElasticSearch (Gormley and Tong, 2015) engine[5] to calculate TF-IDF similarity between full interview texts and the concatenation of the document headline and body, returning the 50 most similar grounding documents for each INTERVIEW conversation. We aim to link documents that would be reasonably relied on by the speakers at the time of the interview, and as such for each interview exclude articles that were published after the interview itself.



(a) (b)

Figure 3: (a) Bar plot depicts test perplexity for linking algorithms: None (no grounding), TF-IDF, and PL/PL3 which indicate probabilistic linking with re-assignment at every 1/3 epochs respectively. Plotting validation perplexity by epoch shows that PL3 converges faster and to a better optimal (b).

Probabilistic Linking While TF-IDF based document linking provides a co-occurence-based similarity measure between documents and conversations, there is no guarantee such linking will improve dialog modeling performance. Thus, we aim to train a linking model such that conditioning on linked documents has a positive effect on dialog modeling performance. We use a two-phase coordinate ascent framework as described in Algorithm 1. In the *Learning* phase, a dialog model is trained based on the available assignments, and its weights are fixed (frozen). Then, in the *Assignment* phase, we compute a re-assignment that maximizes dialog model performance under different possible assignments. Searching over the complete document set is computationally infeasible, so we perform an approximate greedy search over possible documents ordered by their TF-IDF prior score.

We compare the performance of a Transformer (Vaswani et al., 2017) language model provided with grounding documents assigned by different algorithms in Figure 3a. A model without grounding scores by far the worst in terms of perplexity, which indicates that knowledge grounding is important for modeling media dialog. While TF-IDF assignments significantly improve performance compared

[5] https://aws.amazon.com/elasticsearch-service/

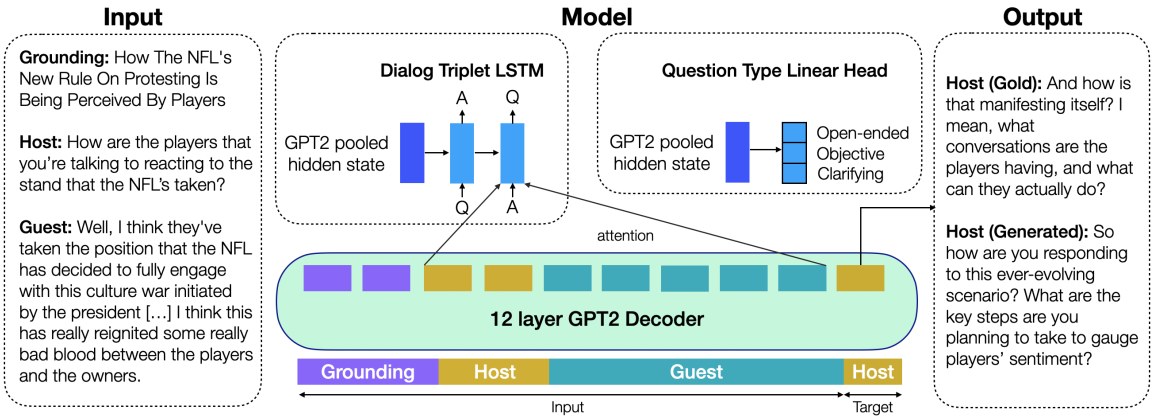


Figure 4: Knowledge grounded generator model with two discourse-specific auxiliary tasks for media dialog

Algorithm 1 Pseudocode for probabilistic linking

```
Initialize document assignments from TF-IDF priors
while average validation perplexity decreases do
    Learning: Update the model with current assignments for N epochs
    for each d in Dialogs do
        Sample K documents from top 50 TF-IDF priors
        for each k in K do
            Condition each response in the dialog with k, and calculate perplexity, aggregate at the dialog level
        end for
        Choose k that yields the lowest perplexity
    end for
    Assignment: Gather all k's for each dialog to update current assignments
end while
```

to no grounding, probabilistic grounding models achieved the best performance. The sudden drops in perplexity values at every third epoch in Figure 3b indicates that the model was well-trained based on current assignments before a new assignments were obtained.

While our articles and conversations come from the same broadcasting source, the NPR interview transcripts generally do not contain links or metadata connecting them with specific grounding documents, and thus there are no ground truth labels available to us. To ascertain that the grounding is relevant, we enlisted two native English speakers who regularly listened to broadcast radio to perform a qualitative evaluation of 100 randomly sampled interview and article pairs. We found that 87% of these pairings are highly relevant, 5% are somewhat relevant and the rest are irrelevant. The inter-annotator agreement measured by Cohen's Kappa was 0.79. The lack of ground truth is something we would argue is not a limitation, rather our probabilistic linking step avoids the dependency on data that is not likely to be available in practice.

5 Modeling Media Dialog

A model's ability to learn underlying discourse dynamics is reflected in its performance on downstream tasks. Here, we assess how well our model learns from dialog structure and question-pattern metadata using utterance generation—a simple predictive task that relies on a holistic understanding of grounding knowledge and a dialog history. This serves as an initial measure of understanding of discourse patterns and grounding even if the exact dialog produced can vary.

We treat knowledge-grounded response generation in the media dialog setting as a language modeling task: given a dialog history H and a grounding knowledge document K, we seek to predict the next utterance x by maximizing the likelihood $p(x|H, K)$. The dialog history is composed of turns spoken by both the interviewer and interviewee where each utterance is provided with the role annotation. We only model interviewer (*host*) responses, which aim to moderate the conversation via questions, follow-ups, and acknowledgements. To understand the effect of dialog structure and question types in response modeling, we introduce two *auxiliary losses* to influence generation—a multi-task setup that has seen success in goal-oriented dialog generation (Luan et al., 2017).

5.1 Knowledge Grounded Generator

We use a common decoder-only model for knowledge-grounded dialog generation (Gopalakrishnan et al., 2019): GPT2 (Radford et al., 2019), a pre-trained Transformer decoder. As model input, we concatenate tokenized grounding documents, dialog history, and the target response. To distinguish each section, we add jointly-learned segment embeddings—{`Grounding`, `Host`, `Guest`}—

Model	Dialog Pattern Pred. Accuracy	Question Type Pred. F1
KGG + Prob. Ground.	38.5	68.8
+ Dialog Pattern	86.3	76.2
+ Question types	**87.9**	**90.5**

Table 3: Performance on auxiliary tasks: Dialog Pattern prediction and Question Type prediction

to each input token. We demonstrate in Section 5.3 that such segment embeddings are essential for this kind of dialog modeling. We only consider target tokens for cross-entropy loss calculation with the conditional likelihood $p(x|H, K)$.

5.2 Predicting Look-ahead Dialog Patterns

Following Section 4.1, we use a generative model to explore the role of response type triplets in structuring media dialog (stemming from an interviewer utterance (Heritage, 1985)). Following response type triplets defined in Section 4.1, we predict the pattern of the dialog triplet beginning with the generated host question as an auxiliary predictive task alongside host utterance generation.

We treat this as a sequence transduction task, employing an LSTM (Hochreiter and Schmidhuber, 1997) decoder with an initial hidden state computed by mean-pooling GPT2 final layer hidden states. Consider $\mathbf{s}_i$ the i-th hidden state from the GPT2 decoder for a length L sequence; now for each hidden state $\mathbf{l}_i$ in the LSTM decoder, we also calculate attention over the GPT2 hidden states, where $\{\mathbf{s}_i\}$ are the keys and values, and $\mathbf{l}_i$ is the query, resulting in an attended vector. We concatenate this attended vector with the LSTM hidden state $\mathbf{l}_i$ and then project it to predict the dialog triplet sequence, maximizing the log-likelihood.

5.3 Predicting Question types

We further explore the impact of question types (dialog acts) via another auxiliary task: multi-label classification for host utterance question types (McLeod et al., 2019). We surmise that accurately predicting question types will help infer question framing and wording, improving generation fidelity. Much like dialog pattern prediction, we use a pooled representation of GPT2 hidden states. We produce a score for each of three question attributes—polarity, combativeness, and subjectivity—via a linear projection and optimize via binary cross-entropy loss.

6 Experiments

In our experiments, we seek to answering the following: 1) Does knowledge grounding help generate more topical host responses? 2) Do our two auxiliary discourse losses improve dialog generation performance? 3) Do human raters find responses generated by our model coherent and fluent? Hyperparameter details are in Appendix §A.

Metrics To measure the fidelity of generated responses, we compute BPE perplexity and BLEU (Papineni et al., 2002) between generated and gold utterances. To assess topical accuracy, we calculate the overlap between noun-phrases and named entities in the generated and gold responses. We are also interested in measuring coherence with respect to the context (i.e., grounding documents and dialog history), calculated via the noun-phrase and named entity overlap between generated responses and context. Furthermore, as news interviews are intended to inform audiences, interviewers must ask questions using specific vocabulary and construction. To assess this, we adopt the Normalized Inverse Document Frequency (See et al., 2019) to measure vocabulary specificity via word rarity. Finally since we focus on generating interrogative host responses, we also calculate the percentage of questions asked in the generated responses as a measure of model inquisitiveness.

6.1 Effect of Knowledge Grounding

To assess the usefulness of explicit grounding documents, we first compare dialog models that use and do not use such documents in Section 5.3. Using segment embeddings to mark utterance bounds improves all measures of fidelity, signifying that this is a useful way to leverage speaker role information in dialog modeling using GPT2. Models that use external grounding knowledge outperform non-grounded models by 1-8 points on almost all metrics, suggesting that such grounding is an important component of host response generation models. To assess the impact of our knowledge grounded generator (KGG) architecture, we compare performance against a strong Memory Network (MemNet) baseline for knowledge grounded dialog generation (Dinan et al., 2019). We confirm our choice of a GPT2-based KGG, as it outperforms Memory Networks in all quality metrics.

Next, we compare the impact of document assignments made via TF-IDF and our probabilistic linking (PL) method. We once again see im-

Model	PPL	BLEU	QR	NPOG	NPOC	NEOG	NEOC	NIDF
No Grounding								
Finetuned (FT) GPT2	28.6	15.4	34.2	0.67	0.57	0.92	0.98	0.105
FT GPT2 + Segment	27.5	17.5	49.9	1.70	1.67	1.56	1.55	0.117
Effect of grounding								
MemNet (2019) + TF-IDF	26.5	17.8	43.8	1.86	1.63	1.51	1.62	0.187
MemNet (2019) + Probabilistic Grounding	25.1	17.7	46.9	1.98	2.31	2.89	3.02	0.197
KGG (TF-IDF)	23.5	18.1	48.5	2.73	3.91	3.01	5.58	0.245
KGG (Probabilistic Grounding)	19.6	19.2	53.6	3.24	4.67	3.44	6.78	0.267
Auxiliary Losses								
+ Dialog Pattern	17.2	**21.0**	56.7	3.52	**6.92**	5.16	**7.85**	0.302
+ Question Types	**15.8**	20.3	**58.9**	**3.67**	6.79	**5.89**	7.79	**0.359**

Table 4: Metrics on generated interviewer responses on test set. NPO/NEO = Noun-phrase/Named entity overlap with context (C) and gold (G); QR = Question rate. NIDF is a measure of specificity (See et al., 2019). QR, NPO, NEO are measured in percentages.

Best Model vs. $\rightarrow$	**No Grounding**		**TFIDF**		**Prob. Grounding**		**+ Dialog Pattern**		**Gold**	
Metric $\downarrow$	win	loss	win	loss	win	loss	win	loss	win	loss
Relevance to Dialog History	**85.1***	9.2	**86.5***	3.3	**69.1***	27.6	**61.0***	22.4	36.7	**47.4***
Relevance to Grounding	**94.5***	4.3	**91.6***	1.9	**82.5***	4.7	**84.5**	9.5	30.1	**55.2***
Fluency	**97.2***	0.8	**87.1***	7.8	**62.1**	10.1	**58.7**	11.2	20.8	**24.6***

Table 5: Pairwise comparison between responses generated by our best model (including both discourse analysis auxiliary tasks) vs. responses generated by other baselines as well as the Gold response. All numbers are in percentages with **bold** indicating the highest. Ties are not shown. Entries with * denote significance with $p < 0.05$ from bootstrap tests on 1000 subsets of size 50.

Grounding (PL)	How The NFL's New Rule On Protesting Is Being Perceived By Players
Context	**HOST**: Host: How are the players that you're talking to reacting to the stand that the NFL's taken? **GUEST**: Well, I think they've taken the position that the NFL has decided to fully engage with this culture war initiated by the president [. . .] I think this has really reignited some really bad blood between the players and the owners. **HOST**:
Gold	And how is that manifesting itself? I mean, what conversations are the players having, and what can they actually do?
KGG (TF-IDF)	Can you tell me more about NFL's new rules?
KGG (Probabilistic Linking)	This are some significant changes in NFL's rules. I think the most effect will be on the players.
+ Dialog Pattern	Okay so let's talk about NFL players. What is your gameplan?
+ Question Types	So how are you responding to this ever-evolving scenario? What are the key steps are you planning to take to gauge players' sentiment?

Table 6: Sample generated response on NFL's new rule. When we add discourse specific losses, the models generate questions that bears more coherence to the context as well as ask clarifying questions.

proved fidelity, mirroring our observations from Section 4.3. Models trained using PL document assignments generate utterances with 19-20% higher noun-phrase and named entity overlap with the gold utterance and context, indicating that PL assignments allow the KGG to more strongly condition on the provided context.

6.2 Effect of Auxiliary Tasks

In this experiment, we investigate how predicting dialog patterns and question types impacts the specificity and fidelity of generated host responses. Each auxiliary loss contributes a significant improvement (1-2 points) in perplexity but affects fidelity and topicality in different ways.

With dialog pattern prediction, we observe that generated responses are more coherent with re-

spect to conversational context, seeing 8% and 48% improvements in noun phrase and named entity overlap with dialog history, respectively. This supports the sociolinguistic observation that the interviewer's choice of utterance (i.e., whether to ask a question, and response content) depends on the discourse structure toward which they aim to guide the conversation (Heritage, 1985). Our results suggest that biasing a dialog model to predict future discourse structure can encourage it to more effectively leverage the past dialog structure (from the conversation history). We confirm in Table 3 that this model can predict look-ahead dialog patterns with 86.3% test-set accuracy. In light of findings that vanilla dialog models may not condition well on conversation context (Sankar et al., 2019), our results suggest one possible direction toward improving contextual language modeling for dialog with inherent structure, such as media dialog.

When we add question-type-prediction loss, we see a significant drop in perplexity and improved fidelity. As expected, by inducing our model to predict the question attributes for the target utterance, our model achieves the highest inquisitiveness (58% question rate). It can also accurately predict question types, with 90.5% macro-averaged test set F1 score. Our results suggest that as the model learns to categorize the interviewer response via specific attributes, it simultaneously learns to generate responses with more specific wording. Table 6 contains representative generations from our best model as well as other baselines, showing that when we add additional discourse specific losses, our model appropriately captures the interviewer's clarifying intent and conversation direction. More generation examples are in Appendix §C.

6.3 Human Evaluation

Automatic evaluation of dialog generation quality is still unreliable (Liu et al., 2016; Novikova et al., 2017), and thus we provide evaluation by human users. We perform pairwise comparisons between responses generated by our best system and those generated by four strong baselines: the best model with no grounding, KGG with TF-IDF, KGG with PL, and KGG with dialog pattern prediction. We also compare against the gold response. Our human evaluation study (details in Appendix §B) measures three aspects of response quality on 100 test examples: 1) How relevant the response is with respect to **dialog history**; 2) How relevant the response is with respect to **grounding documents**; and 3) Whether the generated response is **fluent** English.

We observe in Table 5 that human judges prefer responses generated by our best model (with both discourse analysis auxiliary tasks) to baselines by statistically significant margins in almost every case. This indicates that dialog structure and question types are highly useful for generative modeling in a media dialog setting—specifically news interviews. Human raters also found that despite a significant drop in perplexity when adding the question-type prediction loss, the two versions of discourse-conditioned models had similar fluency, indicating similar language modeling performance. We observe an inter-annotator agreement (Cohen's kappa) of 0.79, 0.92, and 0.73 for relevance to dialog history, grounding documents, and fluency, respectively.

7 Conclusion

In this work, we perform the first large-scale analysis of discourse patterns in media dialog, using a new dataset of 23K annotated news interview transcripts: INTERVIEW. Our results mirror findings from linguistic studies of news interviews (Weizman, 2008; Heritage, 1985). We demonstrate that adding auxiliary tasks for discourse pattern and interrogative type prediction helps model such media dialog. We observe that responses depend heavily on external knowledge, and present a probabilistic framework for linking factual documents with a conversation. While we focus on discourse *pattern* analysis, INTERVIEW also supports analysis of temporal patterns in interviewing, argumentation, and knowledge grounding in long conversations.

Acknowledgements We thank Sudha Rao, Sujoy Paul, Digbalay Bose, and anonymous reviewers for providing valuable feedback on this work. This work is partly supported by NSF Award #1750063. Findings and observations are of the authors only, and do not necessarily reflect the views of the funding agency.

References

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W. Cottrell, and Julian J. McAuley. 2020. Rezero is all you need: Fast convergence at large depth. *CoRR*, abs/2003.04887.

Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker. 2000. The rules behind roles:

Identifying speaker role in radio broadcasts. In *AAAI*.

Monika Bednarek. 2006. *Evaluation in media discourse: Analysis of a newspaper corpus*. A&C Black.

Doug Beeferman, William Brannon, and Deb Roy. 2019. Radiotalk: A large-scale corpus of talk radio transcripts. In *INTERSPEECH*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *ICLR*.

Langzhou Chen, Lori Lamel, Jean-Luc Gauvain, and Gilles Adda. Dynamic language modeling for broadcast news. In *INTERSPEECH*.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2019. Distilling the knowledge of BERT for text generation. *CoRR*, abs/1911.03829.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jordan Cohen. 2007. The gale project: A description and an update. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 237–237. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Teun A van Dijk. 2011. *Discourse and communication: New approaches to the analysis of mass media discourse and communication*, volume 10. Walter de Gruyter.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. 2010. The EPAC corpus: Manual and automatic annotations of conversational speech in french broadcast news. In *LREC*.

Norman Fairclough. 1988. Discourse representation in media discourse. *Sociolinguistics*, 17(2):125–139.

Norman Fairclough and Ruth Wodak. 1997. Critical discourse analysis. *Discourse studies: A multidisciplinary introduction*, 2:258–284.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *EMNLP*.

Kunihiko Fukushima. 1988. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130.

Augusto Gnisci and Marino Bonaiuto. 2003. Grilling politicians: Politicians' answers to questions in television interviews and courtroom examinations. *Journal of language and social psychology*, 22(4):385–413.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Interspeech*.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. "O'Reilly Media, Inc.".

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. In *ICANN*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Pentti Haddington. 2004. Stance taking in news interviews. *SKY Journal of Linguistics*, 17:101–142.

John Heritage. 1985. Analyzing news interviews: Aspects of the production of talk for an'overhearing'audience. *Handbook of Discourse Analysis, vol. III: Discourse and Dialogue*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Qian Huang, Zhu Liu, Aaron E. Rosenberg, David C. Gibbon, and Behzad Shahraray. 1999. Automated generation of news content hierarchy by integrating audio, video, and text information. In *ICASSP*.

Brian Hutchinson, Bin Zhang, and Mari Ostendorf. 2010. Unsupervised broadcast conversation speaker role labeling. In *ICASSP*.

Lauri Karttunen. 1977. Syntax and semantics of questions. *Linguistics and philosophy*, 1(1):3–44.

Antoine Laurent, Nathalie Camelin, and Christian Raymond. 2014. Boosting bonsai trees for efficient features combination: application to speaker role identification. In *INTERSPEECH*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*.

Chia-Wei Liu, Ryan Lowe, Iulian 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 *EMNLP*.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. *CoRR*, abs/1908.03265.

Yang Liu. 2006. Initial study on automatic identification of speaker role in broadcast news speech. In *NAACL-HLT*.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In *IJCNLP*.

Mingyu Derek Ma, Kevin Bowden, JiaQi Wu, Wen Cui, and Marilyn A. Walker. 2019. Implicit discourse relation identification for open-domain dialogues. In *ACL*.

Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison W. Cottrell. 2020. Speech recognition and multi-speaker diarization of long conversations. *CoRR*, abs/2005.08072.

Sarah McLeod, Ivana Kruijff-Korbayova, and Bernd Kiefer. 2019. Multi-task learning of system dialogue act selection for supervised pretraining of goal-oriented dialogue policies. In *SIGDial*. Association for Computational Linguistics.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *CoRR*, abs/1909.08855.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP*.

Paul Placeway, S Chen, Maxine Eskenazi, Uday Jain, Vipul Parikh, Bhiksha Raj, Mosur Ravishankar, Roni Rosenfeld, Kristie Seymore, M Siegler, et al. 1997. The 1996 hub-4 sphinx-3 system. In *Proc. DARPA Speech recognition workshop*, volume 97. Citeseer.

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

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *ACL*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL-HLT*.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *EMNLP*.

Brian G Southwell, Emily A Thorson, and Laura Sheble. 2018. *Misinformation and mass audiences*. University of Texas Press.

Stephanie M Strassel. 2004. Linguistic resources for effective, affordable, reusable speech-to-text. In *LREC*.

Suus MJ Van Hekken and Wim Roelofsen. 1982. More questions than answers: A study of question–answer sequences in a naturalistic setting. *Journal of Child Language*, 9(2):445–460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey. 2011. Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *ICASSP*.

Elda Weizman. 2008. *Positioning in media dialogue: Negotiating roles in the news interview*, volume 3. John Benjamins Publishing.

Stephen J. Wright. 2015. Coordinate descent algorithms. *Math. Program.*, 151(1):3–34.

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. *CoRR*, abs/1911.00536.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

A Implementation Details

Dataset Table 7 provides the statistics for train-dev-test splits on INTERVIEW. We avoid modeling salutations and sign-offs (which tend to be formulaic, and specific to the radio station) by restricting the target turns to those with at least three prior turns and two following turns of conversation, resulting in a target training set of 87K host-only turns and 11K host-only turns for dev and test. We perform BPE tokenization with the GPT2Tokenizer[6].

Network architectures For probabilistic linking, we use a 6-layer encoder-decoder Transformer model (Vaswani et al., 2017). The input to the model consists of grounding document followed by dialog history. The output is the next response in the dialog. To speed up the learning phase, we use ReZero initialization (Bachlechner et al., 2020) that do not require learning weight warm-up schedule. We also observe that performing reassigning at every epoch results in noisy update in assignments and weaker local optima is achieved at the end. When we switch the reassignment phase for every third epoch, the learning stabilizes mirroring a line search (Wright, 2015) from coordinate descent optimization.

For the media dialog generation model, we use GPT2 (Transformer with 12 layers, 768 hidden size, 12 heads, and 117M parameters—`gpt2-small`[7]) as the base architecture. Our best model KGG with two discourse-specific auxiliary losses has 124M parameters.

Hyperparameters We use history size 5 and number of grounding documents as 5. We use the RAdam optimizer (Liu et al., 2019) and the learning rate was set at $6.25e-5$ with a linear decay of step size 10^{-1} per epoch. The loss coefficients in the multi-task loss function for dialog modeling loss, dialog pattern prediction loss and question type prediction loss were 2.0, 1.0, and 1.0 respectively.

Training Each model converged in 3 epochs on an average with batch size 4 in a TITAN X (Pascal) GPU that took 6 hours in total. While training, we only observe perplexity on the validation set to employ an early-stopping criteria.

[6] https://huggingface.co/transformers/model_doc/gpt2.html

[7] https://github.com/huggingface/transfer-learning-conv-ai

Split	# Episodes	# Turns	# Sentences	# Words
Train	18,971	364,461	994,163	17.4 M
Dev	2,371	45,502	123,861	2.2 M
Test	2,372	44,776	122,088	2.1 M
2P	**23,714**	**454,739**	**1,240,112**	**21.7 M**
Full	**105,848**	**3,199,856**	**7,455,662**	**126.7 M**

Table 7: Statistics from two-party (2P) and multi-agent (Full) INTERVIEW dataset

B Evaluation

B.1 Human Evaluation

For human evaluation, we hired two Anglophone (Lifetime HIT acceptance $\% > 80$) annotators for every human-evaluated test generation. Figure 5 shows a sample question for a human judge for the pairwise comparison of a response generated by our best model (KGG with two discourse-specific auxiliary losses) and a response generated by a baseline for three aspects—coherence to dialog history, coherence to grounding, and English language fluency.

C Generation Examples

See Table 8 for a sample dialog history and generated host responses from each of our baseline and our best model—KGG with two auxiliary losses.

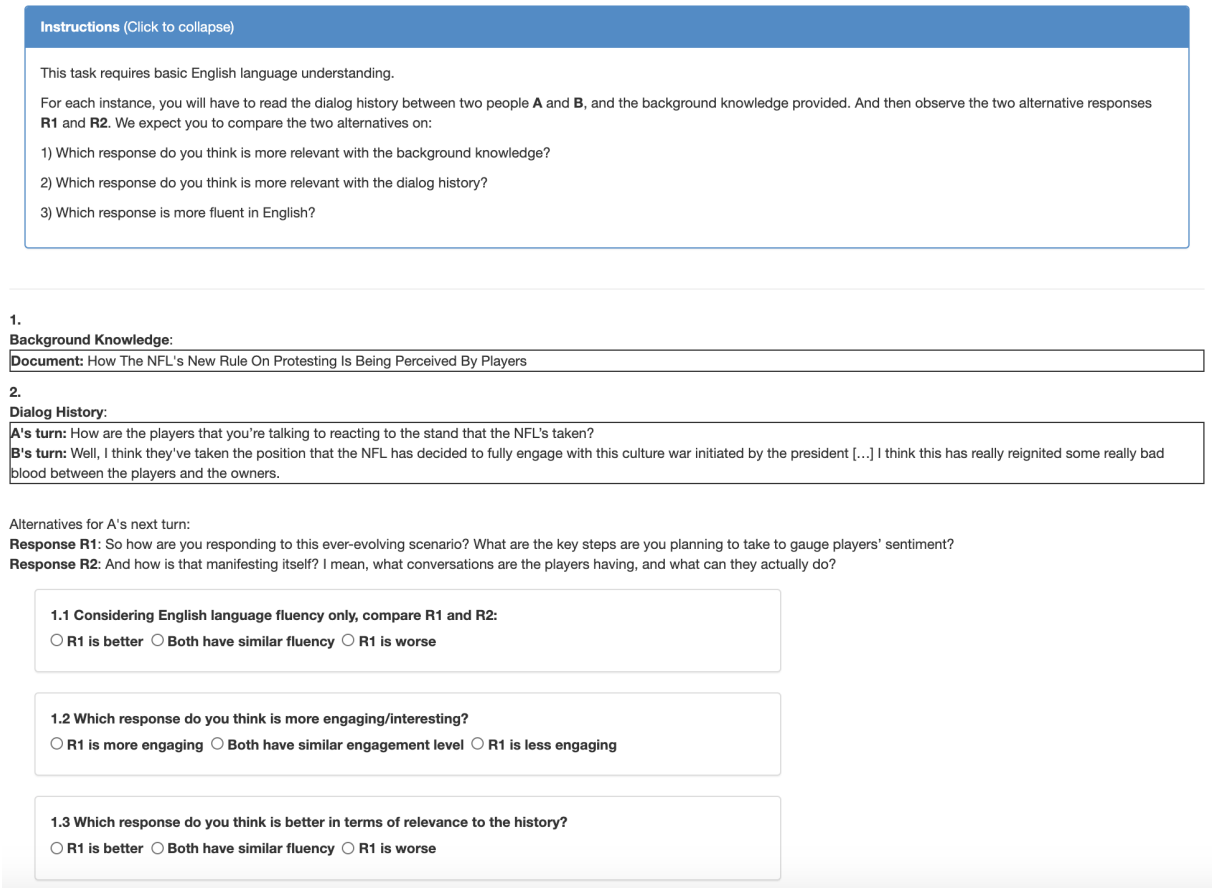


Figure 5: Human evaluation for pairwise comparison between our best model and another baseline.

Grounding (PL)	Trump's Week Of 'Fire And Fury'
Context	**GUEST**: Steve Bannon is quoted as saying [. . .] the president has lost it. Now, are you supporting a president who is incapable of being entrusted with [. . .] nuclear weapons? **GUEST**: Well - one thing I haven't heard yet is Steve Bannon interviewed [. . .] so look, I think the president of the United States has shown he's very, very capable [. . .]. **HOST**:
Gold	Should he be taunting a dictator with nuclear weapons about the size of his nuclear button?
KGG (TF-IDF)	Well, that's what you've been talking about, right?
KGG (Probabilistic Linking)	What do you see as a future?
+ Dialog Pattern	I am worried about his political position now.
+ Question Types	Do you think it's a good idea to confront a nuclear war?

Table 8: Sample generated response on nuclear threat. KGG with discourse specific losses generate more specific and on-topic responses.