NewsDialogues: Towards Proactive News Grounded Conversation

Siheng Li[1]*, Yichun Yin[2], Cheng Yang[1], Wangjie Jiang[1], Yiwei Li[3]
Zesen Cheng[4], Lifeng Shang[2], Xin Jiang[2], Qun Liu[2], Yujiu Yang[1]†
[1]Shenzhen International Graduate School, Tsinghua University
[2]Huawei Noah's Ark Lab, [3]Beijing Institute of Technology, [4]Peking University
lisiheng21@mails.tsinghua.edu.cn
{yinyichun, shang.lifeng, jiang.xin, qun.liu}@huawei.com
yang.yujiu@sz.tsinghua.edu.cn

Abstract

Hot news is one of the most popular topics in daily conversations. However, news grounded conversation has long been stymied by the lack of well-designed task definition and scarce data. In this paper, we propose a novel task, Proactive News Grounded Conversation, in which a dialogue system can proactively lead the conversation based on some key topics of the news. In addition, both information-seeking and chit-chat scenarios are included realistically, where the user may ask a series of questions about the news details or express their opinions and be eager to chat. To further develop this novel task, we collect a human-to-human Chinese dialogue dataset NewsDialogues, which includes 1K conversations with a total of 14.6K utterances and detailed annotations for target topics and knowledge spans. Furthermore, we propose a method named Predict-Generate-Rank, consisting of a generator for grounded knowledge prediction and response generation, and a ranker for the ranking of multiple responses to alleviate the exposure bias. We conduct comprehensive experiments to demonstrate the effectiveness of the proposed method and further present several key findings and challenges to prompt future research.[1]

1 Introduction

News, especially hot news, is widely discussed in daily conversations, enabling people to connect to others and engage with the public issues they encounter in everyday life (Swart et al., 2017). However, due to the lack of well-designed task definition and the scarcity of training data, news grounded conversation has almost been neglected in dialogue system research (Huang et al., 2020; Ni et al., 2021; Thoppilan et al., 2022).

Dataset	Domain	A-p	C-c	I-s
CMU DoG (Zhou et al., 2018)	Film	✗	✔	✔
India Dog (Moghe et al., 2018)	Film	✗	✔	✗
QuAC (Choi et al., 2018)	Wikipedia	✗	✗	✔
CoQA (Reddy et al., 2019)	Multi-Domain	✗	✗	✔
WoW (Dinan et al., 2019)	Wikipedia	✗	✔	✗
doc2dial (Feng et al., 2020)	Service	✗	✗	✔
WikiDialog (Dai et al., 2022)	Wikipedia	✗	✗	✔
InSCIt (Wu et al., 2022)	Wikipedia	✗	✗	✔
NewsDialogues (Ours)	News	✔	✔	✔

Table 1: The differences between NewsDialogues and other document-grounded dialogue datasets. A-p represents the modeling of agent proactivity, C-c and I-s denotes whether the conversations focus on chit-chat scenario and information-seeking scenario respectively.

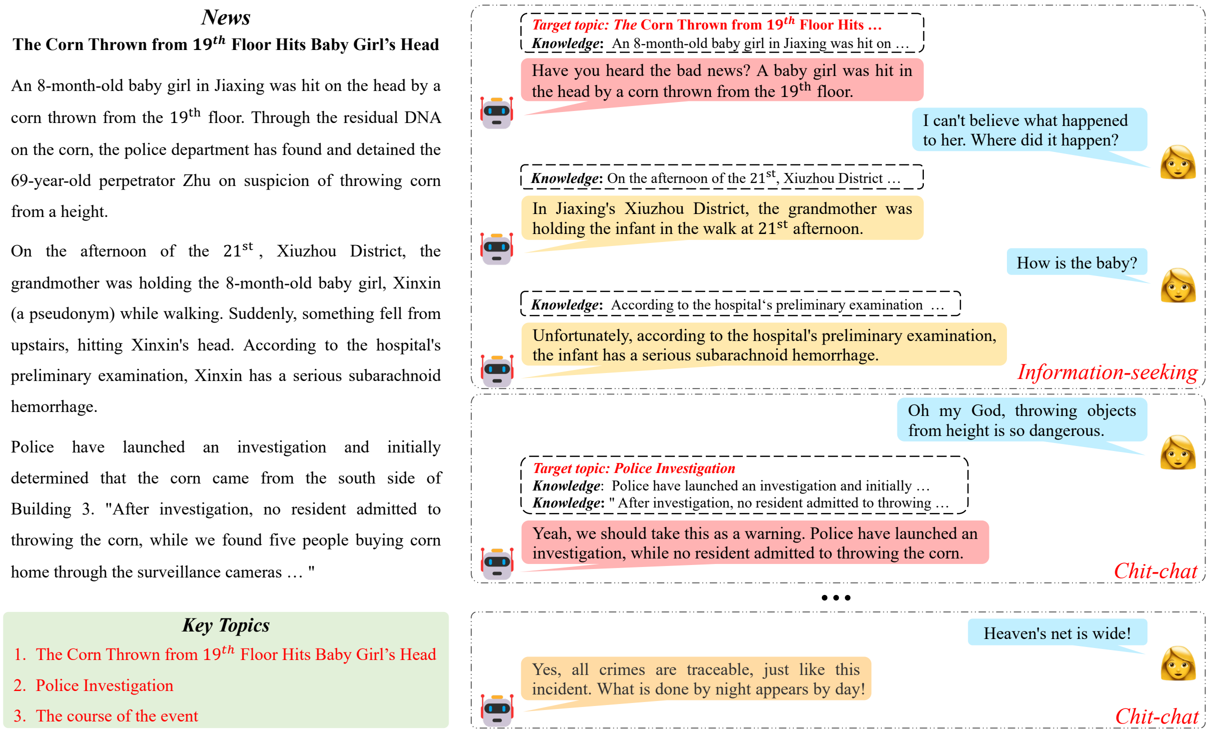
To pursue news grounded conversation, a natural idea is to refer to existing document-grounded conversations. However, there are two major differences. First, as news articles can be long, complex, and time-consuming for human reading, it is important for the dialogue system to be proactive, which means that it can actively introduce news content during the conversation. Therefore, users know more about the news, and the conversations are more interactive and in-depth. However, traditional document-grounded dialogue datasets rarely consider this proactivity explicitly, and the conversations are more user-driven. For example, in *QuAC* (Choi et al., 2018), *doc2dial* (Feng et al., 2020), and *WikiDialog* (Dai et al., 2022), systems mostly respond to user questions passively based on the documents. Second, both chit-chat and information-seeking scenarios (Stede and Schlangen, 2004; Choi et al., 2018) are indispensable for news grounded conversation. Users may ask a series of questions about the news details curiously or express their opinions and be eager to chat. However, existing document-grounded conversations mostly focus on a single scenario of chit-chat or information-seeking rather than both. The work of Choi et al. (2018); Feng et al. (2020); Dai et al. (2022) considers the information-seeking scenario,

[*] This work is done when Siheng Li is an intern at Huawei Noah's Ark Lab.

[†] Corresponding author.

[1] The project repository is available at `https://github.com/SihengLi99/NewsDialogues`.

Findings of the Association for Computational Linguistics: ACL 2023, pages 3634–3649
July 9-14, 2023 ©2023 Association for Computational Linguistics



News

The Corn Thrown from 19th Floor Hits Baby Girl's Head

An 8-month-old baby girl in Jiaxing was hit on the head by a corn thrown from the 19th floor. Through the residual DNA on the corn, the police department has found and detained the 69-year-old perpetrator Zhu on suspicion of throwing corn from a height.

On the afternoon of the 21st , Xiuzhou District, the grandmother was holding the 8-month-old baby girl, Xinxin (a pseudonym) while walking. Suddenly, something fell from upstairs, hitting Xinxin's head. According to the hospital's preliminary examination, Xinxin has a serious subarachnoid hemorrhage.

Police have launched an investigation and initially determined that the corn came from the south side of Building 3. "After investigation, no resident admitted to throwing the corn, while we found five people buying corn home through the surveillance cameras … "

Key Topics

1. The Corn Thrown from 19th Floor Hits Baby Girl's Head
2. Police Investigation
3. The course of the event

Target topic: The **Corn Thrown from 19th Floor Hits …**
Knowledge: An 8-month-old baby girl in Jiaxing was hit on …

Have you heard the bad news? A baby girl was hit in the head by a corn thrown from the 19th floor.

I can't believe what happened to her. Where did it happen?

Knowledge: On the afternoon of the 21st, Xiuzhou District …

In Jiaxing's Xiuzhou District, the grandmother was holding the infant in the walk at 21st afternoon.

How is the baby?

Knowledge: According to the hospital's preliminary examination …

Unfortunately, according to the hospital's preliminary examination, the infant has a serious subarachnoid hemorrhage.

Oh my God, throwing objects from height is so dangerous.

Target topic: Police Investigation
Knowledge: Police have launched an investigation and initially …
Knowledge: " After investigation, no resident admitted to throwing …

Yeah, we should take this as a warning. Police have launched an investigation, while no resident admitted to throwing the corn.

Heaven's net is wide!

Yes, all crimes are traceable, just like this incident. What is done by night appears by day!

Figure 1: An example of NEWSDIALOGUES. We translate the original Chinese dialogue to English version for reading convenience. Notice that some content is omitted as the original version is too long, please refer to the original example in Appendix Figure 3.

where the user repeatedly asks questions and the agent answers based on the documents. Another line of research focuses more on chit-chat scenario (Moghe et al., 2018; Dinan et al., 2019), where participants freely talk about specific topics with knowledge from the documents. For real-world applications, both scenarios should be contained naturally.

To bridge these gaps, we propose a novel task named Proactive News Grounded Conversation, which enables dialogue systems to proactively talk about news with humans in a realistic manner. Furthermore, we collect a human-to-human Chinese dialogue dataset NEWSDIALOGUES, which consists of 1K conversations with 14.6K utterances and rich annotations. We include both information-seeking and chit-chat scenarios realistically, and an example is presented in Figure 1. To explicitly model the proactivity, we first annotate the key topics of the news article to summarize the main content of it. Then, the agent can actively lead the conversation based on these topics, as the 1st and 7th utterances in Figure 1. In addition, we carefully annotate the grounded knowledge of each agent utterance, including the target topic and knowledge spans, for a more informative conversation. The major differences between our NEWSDIALOGUES and other document-grounded dialogue datasets are summarized in Table 1.

To further solve the problem, we propose a simple yet effective method Predict-Generate-Rank, which consists of a generator for grounded knowledge prediction and response generation, and a ranker for the ranking of multiple candidate responses to alleviate the exposure bias problem (Zhang et al., 2019; An et al., 2022). We conduct comprehensive experiments based on the state-of-the-art pre-trained language models and dialogue models. Both automatic and human evaluation indicates that our method has substantial improvements over several baselines on NEWSDIALOGUES. Finally, we analyze the major limitations of current models to facilitate future research.

The main contributions are as follows.

- We propose a novel task named Proactive News Grounded Conversation, aiming to empower dialogue systems to proactively talk about news with humans.

- To further develop this task, we build NEWSDIALOGUES, which consists of 1K dialogues with 14.6K utterances and rich annotations.

- Based on NEWSDIALOGUES, we propose a method named Predict-Generate-Rank and conduct comprehensive experiments. The results have shown the great performance of our method.

2 Related Work

Document-Grounded Conversation. A growing area of research is augmenting dialogue systems with external documents. One line of research focuses on the chit-chat scenario. Zhou et al. (2018); Moghe et al. (2018) propose movie grounded conversation, where two participants talk about movies in depth based on related documents. *Wizard of Wikipedia* (Dinan et al., 2019) introduces more topics for conversations, totally 1,365 from Wikipedia articles. To utilize continually updating knowledge, Komeili et al. (2022) propose *Wizard of the Internet*, where dialogue systems can flexibly search relevant knowledge from the internet.

Another line of research focuses on the information-seeking scenario, where dialogue systems help users gather information through conversations (Choi et al., 2018; Reddy et al., 2019; Campos et al., 2020; Qu et al., 2020). Different from traditional question answering systems, the conversation context empowers dialogue systems to address open-ended and exploratory questions that need discussions to explore in depth (Dai et al., 2022). To pursue more interactive, Feng et al. (2020); Guo et al. (2021); Wu et al. (2022) introduce clarification questions, which means that agents can also ask questions when user queries are defined as under-specified. Though helpful for information-seeking needs, these dialogue systems lack chatting ability.

We propose news grounded conversation, which has been neglected in previous research but is indispensable in our daily conversations. In addition, both chit-chat and information-seeking scenarios are considered realistically.

Proactive Dialogue System. The proactivity of dialogue systems has been an open challenge. Previous researches model proactive topic transitions based on well-designed knowledge graphs (KGs) (Wu et al., 2019; Liu et al., 2020). However, KGs are hard to construct and have limited coverage of real-world knowledge (Razniewski et al., 2016). To explore the topic connections, Sevegnani et al. (2021) propose the one-turn topic transition task and collect a crowdsourced dataset *OTTers*. More recently, Cai et al. (2022) use reinforced self-play to train a teacher bot, which can actively convey knowledge during the conversation. However, they encourage token overlap between the generated responses and the grounded documents rather than proactive topic transition.

We propose proactive dialogue generation based on news articles rather than KGs. Specifically, we aim to empower dialogue systems to lead the conversation based on some key topics of the news. To this end, we build NEWSDIALOGUES, including 1K multi-turn dialogues.

3 Proactive News Grounded Conversation

We propose a novel task named Proactive News Grounded Conversation. As shown in Figure 1, a user converses with an agent based on a given news article in each conversation. The conversation begins with the agent, and during the conversation:

- **User** is curious about the news and eager to chat. They can freely ask questions or express their opinions and feelings.
- **Agent** plays the role of a knowledgeable expert. They not only reply to users in a passive way but also proactively lead the conversations based on the key topics of the news.

Following Choi et al. (2018); Dinan et al. (2019); Kim et al. (2022), we introduce an information-asymmetric setting, where only the agent has access to the news article, and the user is eager to know through the conversation. Therefore, the conversation is more open-ended and exploratory, and the agent is more helpful in real-world applications. Both chit-chat and information-seeking scenarios are contained naturally.

4 NEWSDIALOGUES

To further develop this task, we collect a human-to-human Chinese dialogue dataset NEWSDIALOGUES.

4.1 News Article Collection

We manually collect news articles from Toutiao[2], a famous news website in China. The criteria for selection are: (1) We prefer hot news, and thus humans are more eager to talk about it. To this end, we select news articles from the hot list in Toutiao;

[2] `https://www.toutiao.com/`, we discuss the usage policy in Section 7.

#	Dialog Act	User Utterance	Agent Utterance
1	**Chit-chat**	It is indeed necessary to pay more attention to the elderly.	Yes, after all, we will all grow old. Help the old now, and someone will help us in the future.
2	**Chit-chat**	Well, did the girl say why she went there?	I don't know. Maybe the little girl is naughty and parents truly should take care of their children.
3	**Inform**	What happened in the end? Was he saved?	Yes! He was found by a neighbor in time and saved.
4	**Inform**	Is the old man awake now?	He is still in the ICU, it is not clear how is it going, I hope he can recover soon.
5	**Inform**	He is so talented and loving!	Yeah, what he hopes most is to break the gap and barrier between communities and people in the lockdown.
6	**Guide**	-	***Topic: A police takes a choking girl to hospital.*** Have you heard the news about a police taking a choking girl to hospital? It's so touching!
7	**Guide**	She is a genius! Maybe she can go to the Olympics after the training!	***Topic: Inherits good genes from her mother.*** It is possible! I heard that her mother is a physical education teacher, she inherits the good genes and also develops a habit of exercising.
8	**Guide**	So, why did this guy drive after overdosing?	***Topic: Hidden reactions of driving after overdosing.*** Not mentioned in the news, probably he did not understand the harm of driving after overdosing. People often ignore the adverse reactions, but they are very damaging!
9	**Guide**	I see. Are they from an institution? Why so many people?	***Topic: 7 million yuan are swindled.*** It is a fraud gang with many collaborators! When arrested by the police, they had more than 180 mobile phones and swindled more than 7 million yuan.

Table 2: Examples of different dialog acts of the agent. We highlight some key words of inform, guide and answer for unanswerable question, more details in Section 4.2.2 and 4.2.3. We also present the target topic for guide. For reading convenience, we translate the original Chinese to English and omit the dialog history and knowledge spans.

(2) We only collect news articles that do not rely on image information and leave the multi-modality features for future work.

4.2 Dialogue Collection

In NEWSDIALOGUES, each dialogue derives from a real conversation between two human annotators, one as the user and the other one as the agent. The conversation scenario is based on the task definition in Section 3, and the annotation processes for user and agent annotators are as follows.

4.2.1 User Annotator

Utterance Generation. User annotators freely ask questions or express their opinions and feelings. To further investigate the behavior, we also ask them to annotate the dialog acts (Bunt et al., 2010) of their utterances, which are either **Question** or **Chit-chat**. Here, chit-chat represents the comments or feelings of users, e.g., *He is so talented and loving!*.

4.2.2 Agent Annotator

News Understanding. Before the conversation, the agent annotators carefully read the news articles to understand the overview. Then, we ask them to write the key topics of each news article, typically 2-5 short sentences. They can write key topics in their own words or make appropriate modifications to the section titles of the news articles.

Utterance Generation. During the conversation, the agent annotators choose appropriate dialog acts for each utterance. We introduce three acts, and examples are shown in Table 2.

- **Chit-chat.** Naturally chat with the user without news information.
- **Inform.** Passively respond to the user based on the knowledge from the news article. This act is appropriate when the agent answers user questions or replies to user chit-chat utterances with related news information, as the fourth and fifth examples in Table 2.
- **Guide.** Proactively guide the current conversation based on the key topics and knowledge

from the news. According to our analysis, this act is appropriate under the following scenarios: (1) At the dialogue beginning, as the sixth case in Table 2; (2) The current conversation is relevant to a key topic, and the agent can naturally steer the conversation to the topic, as the seventh example in Table 2; (3) When the user asks an unanswerable question, the agent can lead the conversation to a relevant key topic, as the eighth case in Table 2. Details of unanswerable questions are given below.

Furthermore, we find that almost 10% agent utterances first inform relevant news information and then proactively lead the conversation. We annotate these utterances with the guide act, and an example is the last case in Table 2.

Knowledge Grounding. When the act is inform or guide, the agent annotator can choose appropriate text spans from the news article and use them to craft a natural and informative utterance. We annotate these spans at sentence-level, and each sentence is called a knowledge span. Additionally, we annotate the target topic when the act is guide. These annotations are beneficial for modularized dialogue generation (Zhou et al., 2022; Shuster et al., 2022), which has shown great improvements in knowledge utilization.

4.2.3 Unanswerable Questions

During the annotation process, we find a large number of unanswerable questions, which means that there is no direct answer in the news. This phenomenon is common in realistic information-seeking scenarios, because human questions are open-ended and exploratory. Most existing conversational question answering work simply replies to these questions with No Answer (Choi et al., 2018; Reddy et al., 2019; Adlakha et al., 2022). In this paper, we adopt three strategies in order.

- **Inform Relevant Information.** When there is no direct answer, but providing relevant information possibly fulfills user needs (Wu et al., 2022), as the fourth example in Table 2.
- **Guide Topic Proactively.** When there is no relevant information, but the agent can naturally steer the conversation to a relevant key topic, as the eighth case in Table 2.
- **Chit-chat.** When the above strategies are not suitable under the dialogue context, the agent chats with the user, as the second in Table 2.

Categories	**Statistics**	**Proportion**
News Article		
Total	1000	-
Avg. key topics	3.44	-
Avg. length	1289.67	-
Dialogues		
Total	1000	-
Avg. turns	14.59	-
Avg. length of user utterances	17.44	-
Avg. length of agent utterances	47.28	-
User Dialog Acts		
Chit-chat	2449	35.8%
Question	4398	64.2%
Overall	6847	100.0%
Agent Dialog Acts		
Chit-chat	886	11.4%
Guide	2876	37.1%
Inform	3982	51.4%
Overall	7744	100.0%
Strategies for Unanswerable Questions		
Chit-chat	118	11.2%
Guide Topic Proactively	450	42.6%
Inform Relevant Information	489	46.3%
Overall	1057	100.0%

Table 3: Statistics of NewsDialogues.

4.3 Statistics

The statistics of NewsDialogues are shown in Table 3, and there are several noticeable features. First, the news article is long and brings a new challenge to dialogue system research. Second, as shown by the statistics of user dialog acts, both information-seeking and chit-chat scenarios are common in NewsDialogues. The large proportion of user questions (64.2%) indicates that information-seeking scenario is indispensable for real-world applications. Third, unanswerable questions occupy a large proportion of user questions (1057 of 4398). Therefore, it is important for dialogue systems to address these questions properly.

5 Method

5.1 Task formulation

Each conversation is grounded on a news article $\boldsymbol{n}$ with key topics $\boldsymbol{k}$, and the dialogue system learns to generate a response $\boldsymbol{r}$ based on the dialog history $\boldsymbol{d}$. In addition, it should also predict the grounded knowledge $\boldsymbol{g}$, including both the target topic and the knowledge spans for generation, when needed.

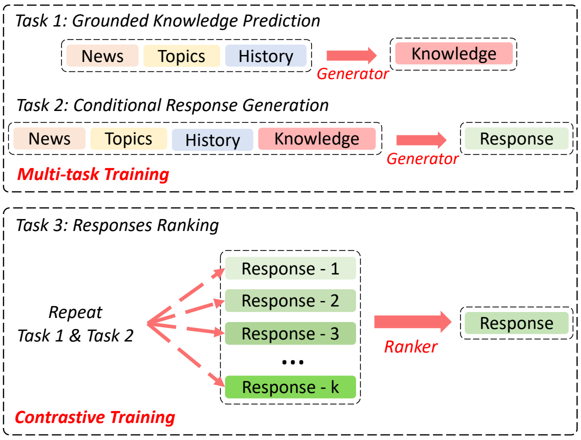


Figure 2: The overview of our Predict-Generate-Rank, including a generator trained with a multi-task objective and a ranker trained with contrastive loss.

5.2 Predict-Generate-Rank

We propose a simple yet effective method, named Predict-Generate-Rank, including a three stage generation process, as shown in Figure 2.

Task 1: Grounded Knowledge Prediction. The model first predicts the grounded knowledge $\boldsymbol{g}$ for response generation. Specifically, we concatenate the target topic and the knowledge spans as $\boldsymbol{g}$[3], and formulate this problem as a task of language generation. The objective is the negative log-likelihood:

$$\mathcal{L}_1 = -\sum_{l=1}^{L} \log P(g_l | g_{<l}, \boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}),$$

where g_l represents the l-th token of $\boldsymbol{g}$, and L is the total length.

Task 2: Conditional Response Generation. Based on the grounded knowledge $\boldsymbol{g}$, our model learns to generate the response autoregressively. The objective function is as follows:

$$\mathcal{L}_2 = -\sum_{t=1}^{T} \log P(r_t | r_{<t}, \boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{g}),$$

where r_t denotes the t-th token of $\boldsymbol{r}$, and T is the total length. We use the ground-truth knowledge for training and the predicted knowledge for inference. Our generator is trained with a multi-task objective: $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, as in Peng et al. (2021).

Task 3: Responses Ranking. One major problem of the above tasks is the gap between the ground-truth knowledge and the predicted knowledge, which results in severe exposure bias (Zhang et al., 2019; An et al., 2022) for text generation. Particularly, the generated response can be low-quality if the predicted knowledge is irrelevant to the dialogue context. To alleviate this problem, we further introduce a ranking task. Specifically, the generator first samples multiple knowledge and generates the responses based on them. Then, a ranker is used to select the best response.

We use a simple strategy to construct datasets for the training of the ranker. First, we finetune the generator on the training set of NewsDialogues, then we use this model to sample knowledge and generate responses on the training set, and we can get $D = \{(\hat{\boldsymbol{g}}_m, \hat{\boldsymbol{r}}_m)\}_{m=1}^{M}$ for each example, where $\hat{\boldsymbol{g}}$ is the predicted knowledge and $\hat{\boldsymbol{r}}$ is the response conditioned on $\hat{\boldsymbol{g}}$. For each $(\hat{\boldsymbol{g}}, \hat{\boldsymbol{r}})$, we compute the matching scores with the ground truth $(\boldsymbol{g}, \boldsymbol{r})$:

$$\begin{aligned}\Delta_1(\boldsymbol{g}, \hat{\boldsymbol{g}}) &= \textbf{Word-Level F1}(\boldsymbol{g}, \hat{\boldsymbol{g}}),\\ \Delta_2(\boldsymbol{r}, \hat{\boldsymbol{r}}) &= \textbf{BLEU}\text{-}4(\boldsymbol{r}, \hat{\boldsymbol{r}}).\end{aligned}$$

The responses with $\Delta_1 > \gamma_1$ and $\Delta_2 > \gamma_2$ are set as positive examples, which means both the knowledge and responses are similar to the ground truths, while other responses are set as negative examples. Then, we can get the training set for the ranker, and the validation set for the ranker is constructed with the same strategy on the validation set of NewsDialogues.

Suppose $\hat{\mathcal{R}}^+$ is the set of positive examples and $\hat{\mathcal{R}}^-$ is the set of negative examples, we train the ranker with contrastive loss:

$$\mathcal{L}_3 = -\sum_{\hat{r}^+ \in \hat{\mathcal{R}}^+} \log \frac{\exp^{s_{\hat{r}^+}}}{\exp^{s_{\hat{r}^+}} + \sum_{\hat{r}^- \in \hat{\mathcal{R}}^-} \exp^{s_{\hat{r}^-}}},$$

where $s_{\hat{r}} = D_\phi([\boldsymbol{d}, \hat{\boldsymbol{r}}])$ is the ranker score and D_ϕ represents the ranker, which is BERT (Devlin et al., 2019) in this paper. The input is the concatenation of the dialogue history $\boldsymbol{d}$ and the response $\hat{\boldsymbol{r}}$, and $s_{\hat{r}} \in \mathcal{R}$ is computed by the representation of `[CLS]`token and a linear projection layer. We pretrain the ranker on DuConv (Wu et al., 2019) and KdConv (Zhou et al., 2020) to better capture the relation between dialogue histories and responses, more details are given in Appendix C.

Inference. For inference, the generator first samples k grounded knowledge and generates re-

[3] Both the target topic and the knowledge spans can be none, depending on the dialog act. When they are none, $\boldsymbol{g}$ is an empty string.

sponses based on them. Then, we use the ranker to select the response with the highest score.

6 Experiments

6.1 Baselines

Dialogue Model. We first investigate the performance of dialogue models. Specifically, we finetune the models on NewsDialogues with only dialogue data, the input is the dialogue history and the target is the ground-truth response. As NewsDialogues is based on Chinese, we evaluate the performance of Chinese dialogue models, CDialGPT (Wang et al., 2020) and EVA2.0 (Gu et al., 2022). EVA2.0 has shown the state-of-the-art performance on Chinese dialogue generation.

End-to-end Model. We finetune the pre-trained language models to predict the grounded knowledge and generate the response based on it sequentially. The training process is the same as our prediction and generation task with a multi-task objective. We evaluate a series of models, including BLOOM (Scao et al., 2022) (Multilingual GPT), mBART (Tang et al., 2020) (Multilingual BART), mT5 (Xue et al., 2020) (Multilingual T5), Chinese GPT (Zhao et al., 2019), Chinese BART (Wang et al., 2022) and Chinese T5 (Wang et al., 2022).

6.2 Implementation

We randomly split NewsDialogues into the train / validation / test sets with a ratio of $8:1:1$, and the numbers of dialogues are 800, 100, and 100. For our Predict-Generate-Rank model, we use Chinese T5 as the generator (Wang et al., 2022) and Mengzi-Bert-base (Chinese BERT) (Zhang et al., 2021) as the ranker. The γ_1 and γ_2 are set as 50 and 15, and the candidate num k is set as 16. More details are shown in Appendix C.

6.3 Automatic Evaluation

Metrics. We adopt BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and Distinct (Li et al., 2016) for the evaluation of response generation. In addition, we compute Topic F1 score to evaluate topic prediction and word-level F1 score for knowledge span prediction (Span F1) as in Choi et al. (2018).

Results. As shown in Table 4, dialogue models perform less competitively than other models. The reason stems from the lack of news information, which is indispensable for NewsDialogues. In addition, dialogue models show the best diversity, and we conjecture this benefits from the pre-training with large-scale conversation data, which contains abundant topics. For end-to-end models, BART performs poorly as it uses absolute position embedding with the maximum length of 1024, which is not sufficient when the news article is long. T5 models with relative position embedding and BLOOM with the maximum length of 2048 can alleviate this problem. The proposed Predict-Generate-Rank improves the performance substantially, except for diversity. We focus more on the relevance between predicted responses and ground-truth responses, which is reflected by other metrics.

6.4 Human Interactive Evaluation

To investigate the performance more realistically, we employ human annotators to converse with different models, humans acting as users while models acting as agents. As human interactive evaluation has a high cost, we only evaluate the best end-to-end model Chinese T5 and our Predict-Generate-Rank. More details are shown in Appendix D.

Metrics. (1) *Fluency*: whether the response is fluent and understandable. (2) *Coherence*: whether the response is coherent and consistent with the context. (3) *Naturalness*: If the response has a target topic, is the topic transition natural and appropriate? (4) *Knowledgeability*: whether the agent is knowledgeable of the news and uses knowledge reasonably. (5) *Proactivity*: whether the agent is proactive and helps you understand the content of the news. (6) *Engagingness*: whether the conversation is engaging and gives you a happy surprise. The first three metrics are utterance-level, while others are dialogue-level. Each score is on a scale from 1 to 3, meaning bad, moderate, and good.

Results. As shown in Table 5, two models show comparable fluency and coherence, and both are far from perfect. For the naturalness of topic transition, Predict-Generate-Rank performs slightly better. Surprisingly, the human score is only 2.60, which indicates the challenge of natural topic transition. Regarding the dialogue-level metrics, our model greatly improves knowledgeability and proactivity, which is consistent with the better performance on topic and knowledge span prediction in automatic evaluation. Furthermore, human evaluators feel more engaged when talking with Predict-Generate-Rank. Nevertheless, there is a large gap between current models and humans in many aspects, indicating plenty of room for improvement.

Model	**Topic F1**	**Span F1**	**BLEU-1**	**BLEU-2**	**BLEU-4**	**ROUGE-2**	**ROUGE-L**	**Distinct-2***
Dialogue Model								
CDial-GPT (Wang et al., 2020)	-	-	14.22	4.56	0.27	2.83	13.32	47.65
EVA2.0 (Gu et al., 2022)	-	-	13.72	3.56	0.14	2.11	13.35	**50.57**
End-to-end Model								
BLOOM (Scao et al., 2022)	41.00	28.42	18.78	9.23	3.23	7.34	19.18	46.47
mBART (Tang et al., 2020)	13.47	28.74	14.33	8.42	4.21	7.83	17.33	37.60
mT5 (Xue et al., 2020)	35.49	27.94	17.18	8.46	3.35	7.06	18.40	47.05
Chinese GPT (Zhao et al., 2019)	37.37	23.96	16.36	7.58	2.28	5.67	16.59	37.50
Chinese BART (Wang et al., 2022)	13.01	26.73	15.78	7.08	1.39	5.82	18.95	39.49
Chinese T5 (Wang et al., 2022)	41.92	39.66	25.42	16.03	8.41	13.92	26.33	45.41
Predict-Generate-Rank	**43.03**	**43.35**	**28.88**	**19.47**	**10.99**	**17.41**	**29.98**	42.45
Human	100.00	100.00	100.00	100.00	100.00	100.00	100.00	51.06

Table 4: Automatic evaluation on the test set of NEWSDIALOGUES. All metrics evaluate the relevance between generations and ground truths except Distinct-2. We list Distinct-2 for the reference of diversity, which is the proportion of distinct bigrams in the total generations and has no relation with the ground truths.

Model	**Flu.**	**Coh.**	**Nat.**	**Kno.**	**Pro.**	**Eng.**
Chinese-T5	2.51	1.94	2.09	1.91	1.66	1.63
Predict-Generate-Rank	**2.53**	**1.99**	**2.16**	**2.15**	**1.92**	**1.71**
Human	2.97	2.91	2.60	2.95	2.80	2.70

Table 5: Human Interactive Evaluation on NEWSDIALOGUES, where Flu., Coh., Nat., Kno., Pro. and Eng. represent *Fluency*, *Coherence*, *Naturalness*, *Knowledgeability*, *Proactivity* and *Engagingness* respectively.

Number	**Topic F1**	**Span F1**	**BLEU-4**	**ROUGE-L**
$k = 1$	41.92	39.66	8.41	26.33
$k = 4$	42.13	43.44	10.33	29.00
$k = 8$	41.86	**43.56**	10.73	29.81
$k = 16$	**43.03**	43.35	**10.99**	**29.98**
$k = 24$	42.61	43.13	10.76	29.17

Table 6: Analysis studies on the candidates number k of Predict-Generate-Rank.

6.5 Impact of Ranking

We conduct experiments to investigate the impact of the ranking task. As shown in Table 6, the performance improves when more candidates are generated, and the Span F1 score has an improvement of 3.78 when only four candidates are generated. Our method gets the best results when $k = 16$, which is the default setting in this paper. According to our manual check, the ranker helps select more relevant responses, thus contributing to the improvements.

6.6 Discussion

Based on the above results, we conclude three major defects of current models. First, they have poor conversation ability, as the low human score in *fluency* and *coherence*. This problem derives from the scale of NEWSDIALOGUES, and a possible way is using the large-scale conversation data in the general domain for pre-training. Second, current models cannot use news knowledge appropriately, as the low Span F1 and *Knowledgeability*. According to our analysis, the reasons are in many aspects: (1) The grounded news is typically long and complex. (2) Many utterances are contextual, and the dialogue system needs to resolve the frequent coreference and information omission (Elgohary et al., 2019) for knowledge extraction. Considering the second utterance in Figure 1, the agent needs to know that “her” represents the “baby girl” in the first utterance. (3) Rather than answering factoid questions in most existing QA datasets, the conversation scenario is much more open-ended, and commonsense reasoning ability is necessary. As the 4th example in Table 2, only when the dialogue system knows the relation between “awake” and “ICU”, can it find the knowledge for a generation. Third, current models are incapable of natural and proactive topic transitions, as the low Topic F1, *Naturalness*, and *Proactivity*. This also stems from the lack of commonsense knowledge and reasoning skills to capture the relations between current topics and relevant topics. This is a valuable characteristic of NEWSDIALOGUES, which is challenging but rewarding for dialogue system research.

7 Conclusion

In this paper, we define a novel task named Proactive News Grounded Conversation, where the dialogue system can proactively lead the conversation based on some topics of the news. In addition, we collect NEWSDIALOGUES with 1K dialogues and rich annotations. Furthermore, we propose Predict-Generate-Rank, which consists of a generator trained with a multi-task objective and a ranker trained with contrastive loss. Comprehensive experiments have been conducted to investigate the performance of current models on NEWSDIALOGUES. We hope that our research will spur the development of dialogue systems that are more proactive and knowledgeable in various scenarios.

Limitations

We acknowledge the following limitations of our work.

Limitations of NEWSDIALOGUES. First, we only collect 1K human-to-human conversations with 14.6K utterances due to the high cost of the annotation process (Section 4.2). This brings difficulties for the learning of news grounded dialogue generation. Second, each conversation in NEWSDIALOGUES is grounded on one news article, which may have limited knowledge for real-world applications. We leave the multi-article grounded setting for future work. Third, as mentioned in Section 4.1, the image information in the news article is neglected in this version, which requires further exploration.

Limitations of Experiments. Large language models (LLM) have shown great few-shot learning ability and generation capacity on various tasks, e.g., GPT-3 (Brown et al., 2020), OPT-175B (Zhang et al., 2022) and BLOOM-176B (Scao et al., 2022) etc. It is important to investigate the performance of LLM on NEWSDIALOGUES, while this has been neglected in this work due to the limited computational resources. In addition, it is also valuable to investigate the performance of ChatGPT[4] on NEWSDIALOGUES, and we leave this for our future work.

Ethics Statement

Private Information

We carefully remove all personal information through the data cleaning process: First, we do not include any account information during the data collecting procedure, which means all the data are anonymous. Second, we clean the potential private information such as emails, ID numbers, phone numbers, etc. in the data to further ensure the privacy.

Offensive Content

We have taken two steps to avoid offensive content in NEWSDIALOGUES. First, we ask the annotators not to speak offensive content during the conversations. Second, we manually check all conversations after data collection and throw away the conversations including offensive content.

Terms of Use

Upon acceptance, we will provide all the codes and the proposed dataset NEWSDIALOGUES including conversations, annotations for knowledge and topics, and corresponding URLs for the News according to the terms of use of Toutiao[5]. NEWSDIALOGUES is only used for facilitating dialogue system research and can not be used for any commercial purposes.

Acknowledgements

This work was partly supported by the National Key Research and Development Program of China (No. 2020YFB1708200) , the "Graph Neural Network Project" of Ping An Technology (Shenzhen) Co., Ltd. and AMiner.Shenzhen SciBrain fund.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Trans. Assoc. Comput. Linguistics*, 10:468–483.

Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2022. Cont: Contrastive neural text generation. *CoRR*, abs/2205.14690.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Kôiti Hasida,

[4] https://openai.com/blog/chatgpt/

[5] https://www.toutiao.com/user_agreement/

Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David R. Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Pengshan Cai, Hui Wan, Fei Liu, Mo Yu, Hong Yu, and Sachindra Joshi. 2022. Learning as conversation: Dialogue systems reinforced for information acquisition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4781–4796. Association for Computational Linguistics.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa - accessing domain-specific faqs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7302–7314. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. 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 5917–5923. Association for Computational Linguistics.

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8118–8128. Association for Computational Linguistics.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, and Minlie Huang. 2022. EVA2.0: investigating open-domain chinese dialogue systems with large-scale pre-training. *CoRR*, abs/2203.09313.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Towards more realistic generation of information-seeking conversations. *CoRR*, abs/2205.12609.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8460–8478. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages

110–119. The Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1036–1049. Association for Computational Linguistics.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2322–2332. Association for Computational Linguistics.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv preprint arXiv:2105.04387*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. SOLOIST: building task bots at scale with transfer learning and machine teaching. *Trans. Assoc. Comput. Linguistics*, 9:907–824.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 539–548. ACM.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Simon Razniewski, Fabian M. Suchanek, and Werner Nutt. 2016. But what do we actually know? In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*, pages 40–44. The Association for Computer Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. Otters: One-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2492–2504. Association for Computational Linguistics.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *CoRR*, abs/2203.13224.

Manfred Stede and David Schlangen. 2004. Information-seeking chat: Dialogues driven by topic-structure. In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*. Citeseer.

Joelle Swart, Chris Peters, and Marcel Broersma. 2017. Repositioning news and public connection in everyday life: A user-oriented perspective on inclusiveness, engagement, relevance, and constructiveness. *Media, culture & society*, 39(6):902–918.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. 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 3794–3804. Association for Computational Linguistics.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2022. INSCIT: information-seeking conversations with mixed-initiative interactions. *CoRR*, abs/2207.00746.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. 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 4334–4343. Association for Computational Linguistics.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: an open-source toolkit for pre-training models. 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 - System Demonstrations*, pages 241–246. Association for Computational Linguistics.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7098–7108. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1237–1252. Association for Computational Linguistics.

Model	MSL	#Params
CDial-GPT (Wang et al., 2020)	512	95.5M
EVA2.0 (Gu et al., 2022)	512	970M
BLOOM (Scao et al., 2022)	2048	560M
mBART (Tang et al., 2020)	1024	610M
mT5 (Xue et al., 2020)	2048	582M
Chinese GPT (Zhao et al., 2019)	1024	102M
Chinese BART (Wang et al., 2022)	1024	759M
Chinese T5 (Wang et al., 2022)	2048	784M
Predict-Generate-Rank (Ours)	2048	784M + 102M

Table 7: The maximum sequence length (MSL) and the parameter number of each model.

A Case Study

For reading convenience, we translate the original Chinese conversation to its English version in Figure 1. Take the original version in Figure 3.

B Annotator Profile

We employ 30 crowdworkers with equally distributed genders for our annotations. They are all native Chinese speakers with ages from 20 to 40 years old. In addition, they are from different regions of China. We pay them a wage above the average in their area. It takes 180,000 ChineseYuan (CNY) for constructing NewsDialogues.

C Implementation Details

All our experiments are based on Transformers[6] (Wolf et al., 2020), DeepSpeed[7] (Rasley et al., 2020) and Pytorch Lightning[8].

General Setting. For both encoder-decoder models and decoder-only models, the input sequence is the concatenation of the news, key topics and the dialogue history, the output sequence is the concatenation of the grounded knowledge and response. We truncate the input sequence according to the maximum sequence length of the model when it uses absolute position embedding. For the T5-based models with relative position embedding, we set the maximum sequence length as 2048. The maximum sequence length and parameters of each model are shown in Table 7. All generative models follow the same hyper-parameter setting. For training, we set the learning rate as $5e-5$, batch size as 32, and use Adam optimizer (Kingma and Ba, 2015) with warmup learning rate schedule, the warmup ratio is 0.1. Each model is trained for 2k gradient steps, and we choose the checkpoint with the lowest perplexity score on the validation set for evaluation. For generation, we use Top-*p* sampling (Holtzman et al., 2020) with p=0.9. We run all experiments three times and report the best results in this paper.

Ours. Our generator is trained with the same hyper-parameter setting as above. For the ranker, the learning rate and batch size are 5e-5 and 64 respectively. The optimizer is the same as that of the generator. We set the maximum gradient steps as 20K for the pretraining stage and 10K for the finetuning stage, the checkpoint with the highest accuracy on the validation set is used for evaluation. After processing DuConv and KdConv, we have 257146 examples for pre-training the ranker, where each example has 1 positive response and 7 negative responses which are randomly sampled from the datasets. We randomly split these examples with a ratio of $4:1$ for the training and validation processes of the pre-training stage. For finetuning the ranker on NewsDialogues, we predict 96 grounded knowledge for each example in the training set of NewsDialogues and generate responses based on them, finally we can get 597504 responses. Then, we construct positive responses and negative responses based on $\gamma_1 = 50$ and $\gamma_2 = 15$, each positive response is paired with 7 negative responses as in the pre-trainig stage. Totally, We can get 35159 examples for the training process of the finetuning stage. Using the same method on the validation set of NewsDialogues, we can get 2854 examples for the validation process of the finetuning stage. Our ranker gets 91.73 accuracy at the pre-training stage and 59.28 accuracy at the finetuning stage.

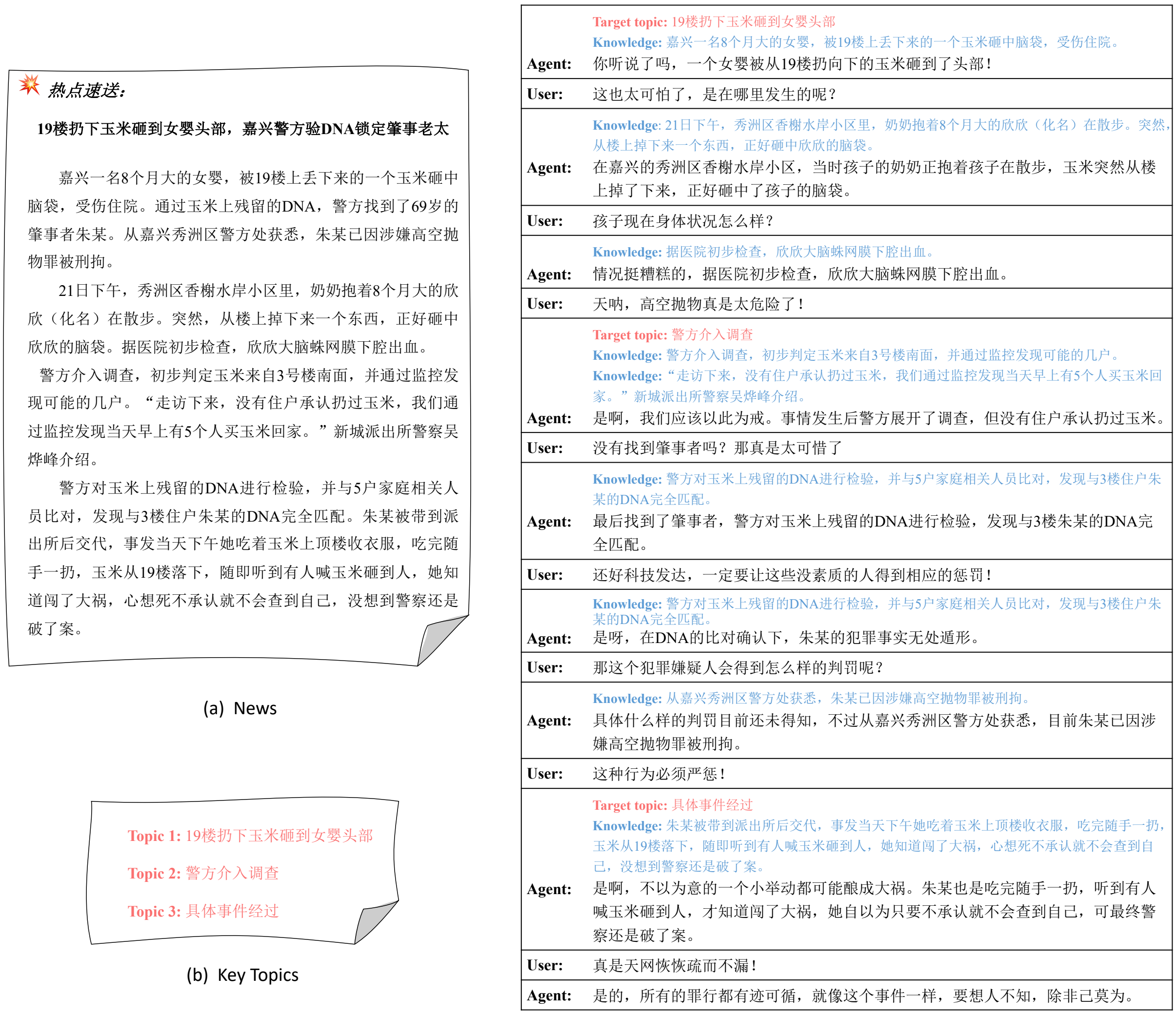
D Human Interactive Evaluation Setting

We employ 4 humans for human interactive evaluation and collect 40 conversations for each model. Specifically, each conversation is grounded on a news article from the test set of NewsDialogues, and contains at least 10 turns, 5 from the human and 5 from the model. In addition, we also select the 40 conversations with the same news articles from the test set to further investigate the performance gap between humans and current models. In total, we have 120 conversations, which are then distributed to 4 human evaluators to score from various aspects.

[6] https://huggingface.co/docs/transformers/index

[7] https://github.com/microsoft/DeepSpeed

[8] https://github.com/Lightning-AI/lightning



热点速送：

19楼扔下玉米砸到女婴头部，嘉兴警方验DNA锁定肇事老太

嘉兴一名8个月大的女婴，被19楼上丢下来的一个玉米砸中脑袋，受伤住院。通过玉米上残留的DNA，警方找到了69岁的肇事者朱某。从嘉兴秀洲区警方处获悉，朱某已因涉嫌高空抛物罪被刑拘。

21日下午，秀洲区香榭水岸小区里，奶奶抱着8个月大的欣欣（化名）在散步。突然，从楼上掉下来一个东西，正好砸中欣欣的脑袋。据医院初步检查，欣欣大脑蛛网膜下腔出血。

警方介入调查，初步判定玉米来自3号楼南面，并通过监控发现可能的几户。“走访下来，没有住户承认扔过玉米，我们通过监控发现当天早上有5个人买玉米回家。”新城派出所警察吴烨峰介绍。

警方对玉米上残留的DNA进行检验，并与5户家庭相关人员比对，发现与3楼住户朱某的DNA完全匹配。朱某被带到派出所后交代，事发当天下午她吃着玉米上顶楼收衣服，吃完随手一扔，玉米从19楼落下，随即听到有人喊玉米砸到人，她知道闯了大祸，心想死不承认就不会查到自己，没想到警察还是破了案。

(a) News

Topic 1: 19楼扔下玉米砸到女婴头部

Topic 2: 警方介入调查

Topic 3: 具体事件经过

(b) Key Topics

Agent:	**Target topic:** 19楼扔下玉米砸到女婴头部 **Knowledge:** 嘉兴一名8个月大的女婴，被19楼上丢下来的一个玉米砸中脑袋，受伤住院。 你听说了吗，一个女婴被从19楼扔向下的玉米砸到了头部！
User:	这也太可怕了，是在哪里发生的呢？
Agent:	**Knowledge:** 21日下午，秀洲区香榭水岸小区里，奶奶抱着8个月大的欣欣（化名）在散步。突然，从楼上掉下来一个东西，正好砸中欣欣的脑袋。 在嘉兴的秀洲区香榭水岸小区，当时孩子的奶奶正抱着孩子在散步，玉米突然从楼上掉了下来，正好砸中了孩子的脑袋。
User:	孩子现在身体状况怎么样？
Agent:	**Knowledge:** 据医院初步检查，欣欣大脑蛛网膜下腔出血。 情况挺糟糕的，据医院初步检查，欣欣大脑蛛网膜下腔出血。
User:	天呐，高空抛物真是太危险了！
Agent:	**Target topic:** 警方介入调查 **Knowledge:** 警方介入调查，初步判定玉米来自3号楼南面，并通过监控发现可能的几户。 **Knowledge:** “走访下来，没有住户承认扔过玉米，我们通过监控发现当天早上有5个人买玉米回家。”新城派出所警察吴烨峰介绍。 是啊，我们应该以此为戒。事情发生后警方展开了调查，但没有住户承认扔过玉米。
User:	没有找到肇事者吗？那真是太可惜了
Agent:	**Knowledge:** 警方对玉米上残留的DNA进行检验，并与5户家庭相关人员比对，发现与3楼住户朱某的DNA完全匹配。 最后找到了肇事者，警方对玉米上残留的DNA进行检验，发现与3楼朱某的DNA完全匹配。
User:	还好科技发达，一定要让这些没素质的人得到相应的惩罚！
Agent:	**Knowledge:** 警方对玉米上残留的DNA进行检验，并与5户家庭相关人员比对，发现与3楼住户朱某的DNA完全匹配。 是呀，在DNA的比对确认下，朱某的犯罪事实无处遁形。
User:	那这个犯罪嫌疑人会得到怎么样的判罚呢？
Agent:	**Knowledge:** 从嘉兴秀洲区警方处获悉，朱某已因涉嫌高空抛物罪被刑拘。 具体什么样的判罚目前还未得知，不过从嘉兴秀洲区警方处获悉，目前朱某已因涉嫌高空抛物罪被刑拘。
User:	这种行为必须严惩！
Agent:	**Target topic:** 具体事件经过 **Knowledge:** 朱某被带到派出所后交代，事发当天下午她吃着玉米上顶楼收衣服，吃完随手一扔，玉米从19楼落下，随即听到有人喊玉米砸到人，她知道闯了大祸，心想死不承认就不会查到自己，没想到警察还是破了案。 是啊，不以为意的一个小举动都可能酿成大祸。朱某也是吃完随手一扔，听到有人喊玉米砸到人，才知道闯了大祸，她自以为只要不承认就不会查到自己，可最终警察还是破了案。
User:	真是天网恢恢疏而不漏！
Agent:	是的，所有的罪行都有迹可循，就像这个事件一样，要想人不知，除非己莫为。

(c) Conversation

Figure 3: An example of NEWSDIALOGUES. For reading conveniently, we translate the original Chinese dialogue to English and omit some information in Figure 1. Here is the original version. During the long conversation, the agent proactively steers the conversation to the key topics of news.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section.

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Ethics Statement.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics Statement.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 and Ethics Statement.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4 and Section 6.

C ☑ Did you run computational experiments?

Section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix C.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix C.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4 and Section 6.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 4.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics Statement.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B.