Towards Proactive News Grounded Conversation

Anonymous ACL submission

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 an average of 14.6 turns and careful annotations for proactive topic transition and grounded knowledge. Furthermore, we introduce two classic methods based on the pre-trained language models to solve this problem, which are the end-to-end method and the read-then-generate method. We conduct extensive experiments to analyze the performance of current models and further present several key findings and challenges to prompt future research. All our code and data will be available after acceptance.

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 scarce data, news grounded conversation has almost been neglected in dialogue system research (Huang et al., 2020; Ni et al., 2021; Thoppilan et al., 2022).

To pursue news grounded conversation, a natural idea is to refer to existing document-grounded conversations. However, there are several differences. First, as news is typically long and complex, it is important for the dialog system to be proactive, which means that it can actively introduce news information related to the dialog context. Therefore, the user can know more about the news, and the conversation is more interactive and in-depth. However, traditional document-grounded datasets rarely consider the proactivity of dialog systems explicitly. Thus the conversations are more user-driven in reality. For example, in *QuAC* (Choi et al., 2018), *doc2dial* (Feng et al., 2020), and *WikiDialog* (Dai et al., 2022), the agent mostly responds 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 conversation research mostly focuses on a single scenario of chit-chat or information-seeking scenario, rather than both. The work of Choi et al. (2018); Feng et al. (2020); Dai et al. (2022) considers the information-seeking scenario, where the user repeatedly asks questions and the agent answers them based on the documents. Another line of research focuses more on chit-chat scenario (Zhou et al., 2018; Dinan et al., 2019; Komeili et al., 2022), where participants talk about specific topics with knowledge from the documents. Compared to the information-seeking scenario, the chit-chat scenario is more casual, in which the user can freely talk about their opinions and feelings.

To bridge these gaps, we propose a new task named Proactive News Grounded Conversation, and collect a human-to-human Chinese dialogue dataset NEWSDIALOGUES, which consists of 1K conversations with an average of 14.6 turns and rich annotations. We include both information-seeking and chit-chat scenarios for a more realistic application, and an example is presented in Figure 1. To explicitly model the proactivity of the dialog system, we first annotate the key topics of each news,

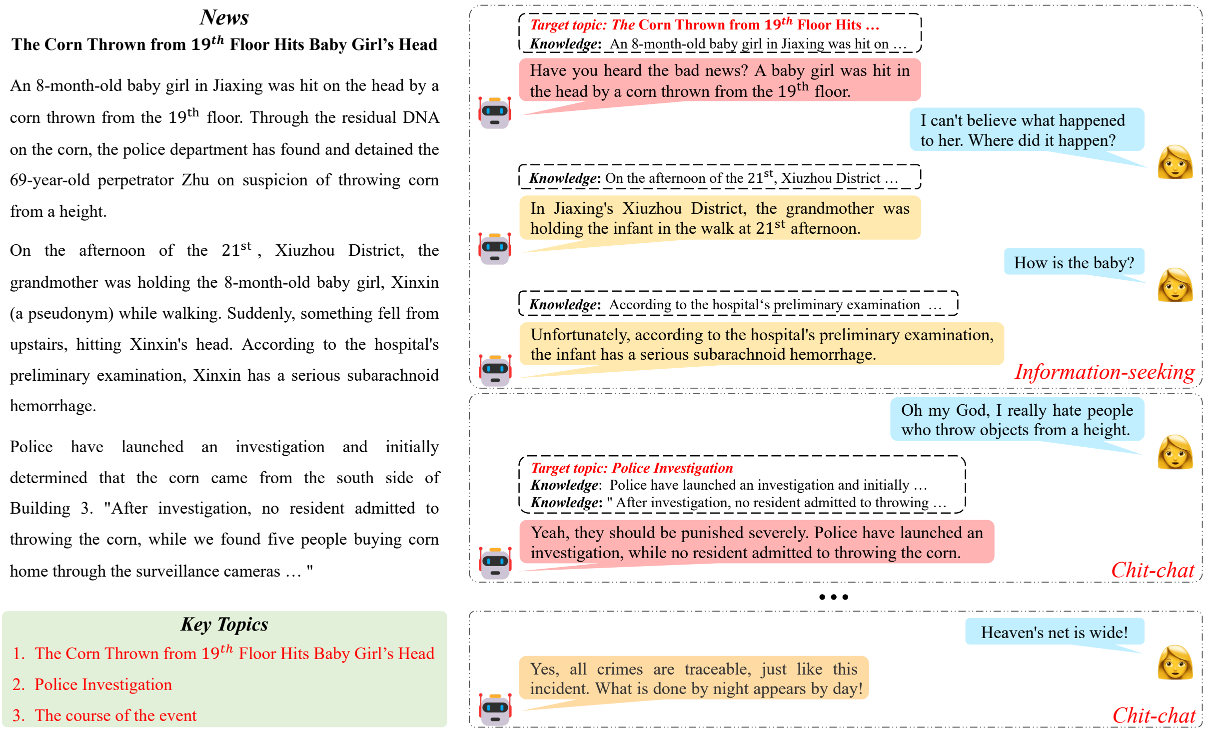

Figure 1: An example of NEWSDIALOGUES. We translate the original Chinese dialogue to English version for reading convenience. Notice that some content is omitted as the original version is too long, please refer to the original example in Appendix Figure 3.

which summarize the main content of it. Then, the dialog system can actively lead the conversation to relevant key topics, as the 1st and 4th utterances of the agent in Figure 1. Thus the dialogue is more in-depth and informative. We carefully annotate whether conduct topic guidance under the dialog context and the target topic if appropriate, more details in Section 4.2. In addition, we annotate the grounded knowledge for each agent utterance at sentence-level for a more informative conversation.

To further solve the problem, we introduce two methods: (1) End-to-end: uses a single language model to generate all text, including the target topic, knowledge spans, and response. (2) Read-then-generate: first predicts the target topic and knowledge spans at a reading stage, then generates a response based on them. We conduct extensive experiments based on these methods, and the state-of-the-art pre-trained language models and dialog models. Results indicate that the read-then-generate method with pre-trained language models performs better in NEWSDIALOGUES. Finally, we analyze the major limitations to facilitate future research.

The main contributions are as follows.

- We propose a novel task named Proactive News Grounded Conversation, aiming to empower dialog systems with more proactivity in news grounded conversation.
- To further develop this task, we build a human-to-human Chinese dialog dataset NEWSDIALOGUES, which consists of 1K dialogs with an average of 14.6 turns and rich annotations.
- Based on NEWSDIALOGUES, we introduce two methods, conduct comprehensive experiments, and provide several key findings.

2 Related Work

Document-Grounded Conversation. A growing area of research is that of 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 the dialogue system can flexibly search documents from the internet.

Another line of research focuses on information-seeking (goal-oriented) scenario. Conversational question answering (Choi et al., 2018; Reddy et al., 2019; Campos et al., 2020; Qu et al., 2020; Anantha et al., 2021) aims to help users gather information through conversations, which is important for addressing more open questions that need discussions to explore in depth (Dai et al., 2022). Furthermore, Feng et al. (2020); Wu et al. (2022) introduce clarification questions, which means the agent can also ask questions when the user query is defined as under-specified. Though helpful, these dialog systems lack chatting ability.

We propose news grounded conversation, which is neglected in previous research but indispensable in our daily life. Both chit-chat and information-seeking scenarios are considered realistically.

Proactive Dialogue System. The proactivity of dialogue system has been an open challenge. Previous work proposes to 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). Sevegnani et al. (2021) propose one-turn topic transition task and collect the dataset *OTTers*. More recently, Cai et al. (2022) propose to actively help users gain knowledge during the conversation. However, they simply encourage token overlap between the generated responses and documents, rather than proactive topic transition.

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

3 Proactive News Grounded Conversation

We propose a new task named *Proactive News Grounded Conversation*. Specifically, a user converses with an agent based on given news, as shown in Figure 1. Each conversation begins with the agent. 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 passively chat with users, but also proactively lead conversations to some key topics of the news.

Following Choi et al. (2018); Kim et al. (2022), we introduce an information-asymmetric setting, which means only the agent has access to the news, the user has not seen the news and is eager to know it from the conversation. Therefore, the conversation is more open-ended and exploratory (Choi et al., 2018), and the agent is more helpful in the application. Furthermore, we do not constrain the content and style of the conversation. Thus it contains both chit-chat and information-seeking scenarios realistically.

4 NEWSDIALOGUES

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

4.1 News Collection

We manually collect hot news from Toutiao[1], a famous news website in China. The criteria for news selection are: (1). We prefer hot news, making humans more eager to talk about it. To this end, we select news from the hot list in Toutiao. (2). We only collect news that does not rely on picture information and leave the multi-model features for future work.

4.2 Dialogue Collection

In NEWSDIALOGUES, each dialogue derives from a real conversation between two human annotators, one as a user and the other one as an 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

Dialogue Generation. User annotators freely ask questions or express opinions and feelings. To further investigate their 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 read the news carefully to understand the overview. Then, we ask them to write

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

Dialog Act	User Utterance	Agent Utterance
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.
Chit-chat	That's fine. 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.
Inform	What happened in the end? Was he saved?	Yes! He was found by a neighbor in time and saved.
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.
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.
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!
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.
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!
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 1: 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.

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 on the section titles of news.

Dialogue Generation. During the conversation, the agent annotators choose appropriate dialog acts for each turn. We introduce three acts, and examples are shown in Table 1.

- **Chit-chat.** Naturally chat with the user without news information.
- **Inform.** Passively respond to the user based on knowledge from the news. Typically when the agent answers user questions or replies to the user chit-chat utterances with related news information, as the fifth example in Table 1.
- **Guide.** Proactively guide the current conversation based on key topics and knowledge from the news. According to our analysis, this action is appropriate under the following scenarios: (1) At the dialogue beginning, as the sixth case in Table 1. (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 1. (3) When the user asks an unanswerable question, the agent can guide the conversation to a relevant key topic, as the eighth case in Table 1. More details of unanswerable questions are given in Section 4.2.3.

Furthermore, we find that almost 10% agent utterances first passively inform relevant news information and then proactively lead the conversation. We also annotate these cases as the guide action, as the last example in Table 1.

Knowledge Grounding. When the act is inform or guide, we annotate the grounded knowledge at sentence-level, 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 have shown improvement in knowledge utilization. We ask them not simply to parrot news text, but to depend on it to craft a natural reply, where oralization and summarization are necessary.

Categories	Statistics	Proportion
News		
Total	1000	-
Avg. key topics	3.44	-
Avg. length	1289.67	-
Dialogs		
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 2: Statistics of NewsDialogues.

4.2.3 Unanswerable Questions

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

- **Inform Relevant Information.** When there is no direct answer, but providing relevant information possibly fulfill user needs (Wu et al., 2022), as the fourth example in Table 1.

- **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 1.
- **Chit-chat.** When the above strategies are not suitable under the dialogue context, the agent chats with the user, as the second in Table 1.

4.3 Statistics

The statistics of NewsDialogues are shown in Table 2, there are several noticeable features. First, understanding the long news brings a new challenge to dialogue system research. Second, 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 in realistic conversation. Third, unanswerable questions occupy a large proportion of user questions (1057 of 4398). Therefore, it is important for dialog systems to address these questions properly.

5 Method

5.1 Task formulation

Each conversation is grounded on news $\boldsymbol{n}$ 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 predict the target topic $\boldsymbol{t}$ and extract knowledge spans $\boldsymbol{s}$ from the news for generation when needed. We introduce two classical methods: (1) End-to-end: uses a single language model to generate all text, including the target topic, knowledge spans, and response. (2) Read-then-generate: first, predict the target topic and knowledge spans, then generate the response based on them.

5.2 End-to-end Method

Thanks to the transferability of pre-trained language models (e.g., GPT, T5), end-to-end methods have shown great progress in dialog generation (Wolf et al., 2019; Hosseini-Asl et al., 2020). Inspired by this, we formulate the problem as a task of language generation and minimize the negative log likelihood of generating string $\boldsymbol{g}$:

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

where $\boldsymbol{g}$ represents the whole generation sequence, including topic, knowledge, and response, as the generation of the end-to-end method in Figure 2. g_l denotes the l-th token, and L is the total length.

5.3 Read-then-generate Method

This method consists of a read stage for topic prediction and knowledge span extraction, and a generate stage for response generation.

Read Stage. We formulate this stage as sentence classification as in extractive summarization[2]

[2]We also try the span selection method as extractive question answering (Rajpurkar et al., 2016), while we find that performance is inferior.

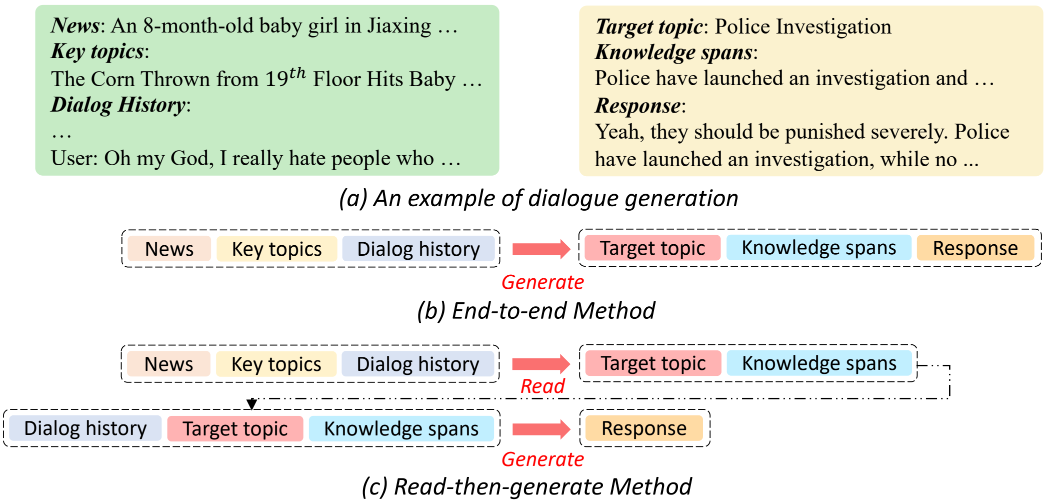

Figure 2: The overview of our methods. *(b)* and *(c)* describe the input and output format of the end-to-end method and the read-then-generate method respectively.

(Liu and Lapata, 2019). As in Figure 2, the input is the dialog history, key topics, and news. In addition, we prepend `[CLS]` for each sentence, including all topics and news sentences. Then, `[CLS]` representation with a binary classification head is used for classification. The objective function is binary cross-entropy, and we adopt a positive weight to alleviate the unbalance problem of positive and negative examples.

$$\mathcal{L}_2 = -\sum_{i=1}^{I}[w \cdot y_i \cdot \log P(x_i|\boldsymbol{n},\boldsymbol{k},\boldsymbol{d}) + (1-y_i) \cdot \log(1 - P(x_i|\boldsymbol{n},\boldsymbol{k},\boldsymbol{d}))],$$

where I denotes the number of sentences, $w \in \mathbb{R}$ is the positive weight, $y_i \in \{0, 1\}$ is 1 if the i-th sentence is selected as target topic or knowledge span, and x_i is the i-th sentence. In the inference time, we select sentences with probability larger than a manually set threshold $\gamma \in (0, 1)$. To process the long news, we choose a bi-directional pre-trained language model, Longformer (Beltagy et al., 2020) for this stage, which is pre-trained with masked language modeling on long documents and supports up to 4096 tokens.

Generate Stage. Based on the predicted topic $\boldsymbol{t}$ and knowledge spans $\boldsymbol{s}$, this stage is used for generating response $\boldsymbol{r}$, as illustrated in Figure 2. Compared to the end-to-end method, it does not need to process long news and thus is more efficient. The objective function is as follows:

$$\mathcal{L}_3 = -\sum_{l=1}^{L}\log P(r_l|r_{<l},\boldsymbol{d},\boldsymbol{t},\boldsymbol{s}).$$

We use the ground-truth topic and knowledge span for training and the predicted topic and knowledge span of read stage at the inference time.

6 Experiments

We conduct a series of experiments to investigate this new task. First, we compare the performance of the introduced two methods built on common pre-trained language models with automatic evaluation. Second, we conduct a human interactive evaluation to further evaluate the methods realistically. Third, we conduct an ablation study to analyze the importance of knowledge grounding, including topic prediction and knowledge span extraction. Finally, we discuss the main limitations of current models in NEWSDIALOGUES.

6.1 Implementation

We randomly split NEWSDIALOGUES into the train / validation / test sets with an ratio of $8:1:1$, the number of dialogues are 800, 100 and 100.

Generation Model. Both the end-to-end and read-then-generate methods are built with the generation models, which are described as follows:

BlOOM (BigScience, 2022). A large multilingual language model with GPT-like decoder-only architecture, we use the 560M parameters version[3].

mT5 (Xue et al., 2020). A multilingual variant of T5 (Raffel et al., 2020), we use the base version with 580M parameters[4].

[3]https://huggingface.co/bigscience/bloom-560m

[4]https://huggingface.co/google/mt5-base

Model	Topic F1	Span F1	BLEU-2	BLEU-4	ROUGE-2	ROUGE-L	Distinct-2	Speedup
				End-to-end				
EVA2.0	0.16±0.3	14.95±1.3	3.50±0.1	0.24±0.0	2.35±0.1	14.00±0.5	30.70±1.2	1.00×
BLOOM	48.67±2.7	37.60±1.6	14.95±0.5	7.21±0.5	13.06±0.6	25.68±0.7	**42.83±1.9**	1.35×
mT5	13.46±1.3	27.10±0.5	8.92±0.1	3.13±0.2	7.21±0.2	18.88±0.3	37.68±0.8	1.74×
				Read-then-generate				
r-EVA2.0	58.60±0.5	43.11±1.3	5.59±0.1	0.51±0.0	3.72±0.2	16.67±0.2	33.11±1.6	1.89×
r-BLOOM	58.60±0.5	43.11±1.3	15.87±0.5	7.93±0.5	13.96±0.5	27.50±0.2	39.31±0.7	3.31×
r-mT5	**58.60±0.5***	**43.11±1.3***	**17.65±0.1**	**10.17±0.1**	**16.29±0.1**	**28.98±0.1**	42.22±0.2	**4.18×**
Human	100	100	100	100	100	100	51.06	-

Table 3: Automatic evaluation on NEWSDIALOGUES, *r*- represents the read-then-generate methods. We report the averages across 4 random seeds, with standard deviations as subscripts. *: The read-then-generate methods use the same model in the read stage, thus have the same Topic F1 and Span F1. Speedup is in terms of the EVA2.0 inference speed and evaluated on the test set with one Tesla V100 32GB GPU and batch size 1. For the read-then-generate method, the inference time contains both the time of read stage and the time of generate stage.

EVA2.0 (Gu et al., 2022). The state-of-the-art open source Chinese dialogue model, we use the large version with 970M parameters[5].

Read Model. We use the Chinese version Longformer[6] (Beltagy et al., 2020) with 330M parameters, which is pre-trained by Wang et al. (2022) with MLM loss.

More implementation details are in Appendix C.

6.2 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 extraction (Span F1) as in Choi et al. (2018).

Results. As in Table 3, EVA2.0 performs much worse than the pre-trained language models in NEWSDIALOGUES, although it has shown state-of-the-art performance in open-domain conversation (Gu et al., 2022). One of the reasons is that pre-trained language models learn more knowledge from the massive unsupervised text across various domains, and facilitates stronger ability in news understanding, thus better performance in news grounded conversation, which is similar with the observation in Zheng et al. (2022). Another reason is that the maximum sequence length of EVA2.0 in the pre-training stage is only 128, which is not sufficient for NEWSDIALOGUES. Therefore, it is a challenge for dialog models to learn knowledge grounded generation when the news text is long and complex, which is indispensable for the information-seeking scenario.

As in Table 3, the end-to-end models perform poorly for topic prediction and knowledge span extraction, resulting in inferior response quality. By the read stage, the read-then-generate method achieves better performance in both topic prediction and span extraction, resulting in better response. We conjecture that one reason is that the sentence classification task more explicitly models the extraction process at the sentence-level rather than the token-level as in language modeling. In addition, read-then-generate models are more efficient, as they do not need to generate the knowledge spans autoregressively.

6.3 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 is high cost, we only evaluate the best end-to-end model BLOOM and the best read-then-generate model *r*-mT5. More details are in Appendix D.

Metrics. (1) *Fluency*: whether the response is fluent and understandable. (2) *Coherence*: whether the response is coherent and consistent with dialogue 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 key content of the news. (6) *Engagingness*: whether the conver-

[5] https://huggingface.co/thu-coai/EVA2.0-large

[6] https://huggingface.co/IDEA-CCNL/Erlangshen-Longformer-330M

Model	Flu.	Coh.	Nat.	Kno.	Pro.	Eng.
BLOOM	2.45	1.93	2.09	1.80	1.60	1.60
r-mT5	2.47	1.94	**2.13**	**2.11**	**1.94**	**1.67**
Human	2.97	2.91	2.60	2.95	2.80	2.70

Table 4: Human Interactive Evaluation on NEWSDIALOGUES, where Flu., Coh., Nat., Kno., Pro. and Eng. represent Fluency, Coherence, Naturalness, Knowledgeability, Proactivity and Engagingness respectively.

Model	BLEU-4	ROUGE-L	Distinct-2
r-mT5	**10.17**±0.1	**28.98**±0.1	**42.22**±0.2
w/o span	5.70±0.2	26.55±0.2	34.87±0.5
w/o topic	7.72±0.1	25.69±0.3	41.33±0.2
w/o both	0.90±0.1	17.27±0.1	30.97±0.6
w/ oracle	22.18±0.2	44.32±0.2	46.82±0.4

Table 5: Ablation Studies on NEWSDIALOGUES. All experiments are performed 4 runs with different random seeds. *w/o* means without, *both* represents span and topic, and *w/ oracle* means with oracle span and topic.

sation is engaging and gives you a happy surprise. The first three metrics are utterance-level, while others are dialog-level. Each score is on a scale from $1-3$, meaning bad, moderate, and good.

Results. As in Table 4, BLOOM and *r*-mT5 show comparable fluency and coherence, and both are far from perfect. For the naturalness of topic transition, *r*-mT5 performs slightly better. Surprisingly, the human score is only 2.60, which shows the challenge of natural topic transition. Regarding the dialog-level metrics, *r*-mT5 greatly improves the knowledgeability and proactivity, which is consistent with the better performance of topic prediction and knowledge span extraction in automatic evaluation. Furthermore, human evaluators feel more engaged when talking with *r*-mT5. In summary, there is still a large gap between current models and humans in many aspects, indicating plenty of room for improvement.

6.4 Ablation Study

We further analyze the importance of knowledge grounding, as in Table 5. The response generation metrics drop largely for both relevance and diversity, when each part is removed. This proves that knowledge grounding is necessary and also indicates that models can learn the knowledge grounding ability with NEWSDIALOGUES. In addition, we also investigate the performance of an oracle model with ground-truth knowledge span and target topic. The large gap shows that there is still great potential for improvement and a promising way is to improve the performance of the read stage.

6.5 Discussion

Based on the above results, we conclude three major defects of current models. First, these models 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 dialog 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 is necessary. As the 4th example in Table 1, only when the dialog systems know the relation between "awake" and "ICU", can they 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 reasoning ability to capture the relations between the current topic and other topics. This is a unique characteristic of NEWSDIALOGUES, which is challenging but rewarding for dialog system research.

7 Conclusion

In this paper, we define a novel task named Proactive News Grounded Conversation, where both chit-chat and information-seeking scenarios are included realistically, and the dialog system can proactively lead the conversation based on some key topics of the news. In addition, we collect NEWSDIALOGUES with 1K dialogues and rich annotations. To further solve the problem, we introduce two classical methods and conduct comprehensive experiments and analyses. We hope that our research will spur the development of dialog systems that are more proactive and knowledgeable in various conversation scenarios.

Ethical Considerations

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[7]. NEWSDIALOGUES is only used for facilitating dialogue system research and can not be used for any commercial purposes.

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.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 520–534. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

BigScience. 2022. Bigscience language open-science open-access multilingual (bloom) language model. *International*.

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

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

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.

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.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

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)*, pages 8460–8478, Dublin, Ireland. 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*, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3728–3738. 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*, pages 2322–2332, Brussels, Belgium. 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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

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.

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.

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.

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.

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

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*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *CoRR*, abs/2202.13047.

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.

A Case Study

For reading convenience, we translate the original Chinese conversation to its English version. Take an example 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 have various occupations and are from different regions of China. We pay them a wage above the average in their area.

C Implementation Details

All our experiments are based on Transformers[8] (Wolf et al., 2020), DeepSpeed[9] (Rasley et al., 2020) and Pytorch Lightning[10].

Generation Setting. For the end-to-end method, the maximum sequence length is 2048 for BLOOM and mT5, and 512 for EVA2.0 which is its maximum supportable length. For the read-then-generate method, the maximum sequence length is 512 for all models. All generative models follow the same hyper-parameter setting. For training, we set the learning rate as $3e-5$, 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-k and Top-p sampling (Holtzman et al., 2020) with k=30, p=0.9 and temperature=0.7.

Read Stage Setting. The maximum sequence length for Longformer at the read stage is set as 4096 thanks to the sparse attention pattern. For training, the learning rate, batch size, gradient steps and positive weight are $5e-5$, 32, 3k, and 15 respectively, and the optimizer is the same as the generation setting. We choose the checkpoint with the best combinational F1 score (Topic F1 + Span F1) on the validation set for evaluation, the threshold γ is 0.5.

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

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

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

D Human Interactive Evaluation Setting

We collect 40 conversations with 4 humans for each model, where the news comes from our test set. Each conversation contains at least 10 turns, 5 from the human and 5 from the model. In addition, we also select 40 conversations from test dataset with the same news 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.

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，这些随意高空抛物的人真的太没素质了！
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，他们应该受到严惩。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 in NEWSDIALOGUES. During the long conversation, the agent proactively steers the conversation to the key topics of news.