A Benchmark for Understanding and Generating Dialogue between Characters in Stories

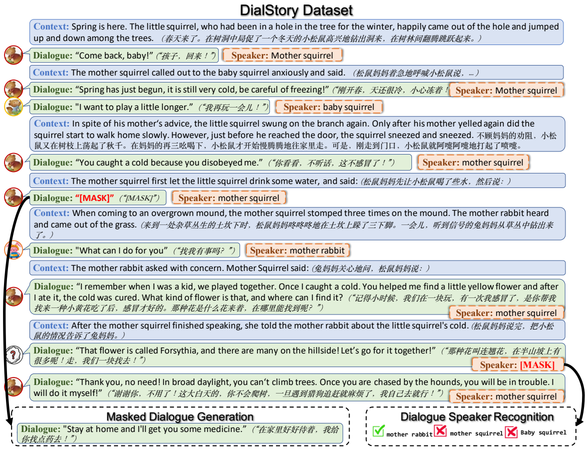
Jianzhu Yao, Ziqi Liu, Jian Guan, Minlie Huang[*]
The CoAI group, Tsinghua University, Beijing, China
Department of Computer Science and Technology, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology
{yjz19, liuzq19, j-guan19}@mails.tsinghua.edu.cn;
aihuang@tsinghua.edu.cn

Abstract

Many classical fairy tales, fictions, and screenplays leverage dialogue to advance story plots and establish characters. We present the first study to explore whether machines can understand and generate dialogue in stories, which require capturing traits of different characters and the relationships between them. To this end, we propose two new tasks including Masked Dialogue Generation and Dialogue Speaker Recognition, i.e., generating missing dialogue turns and predicting speakers for specified dialogue turns, respectively. We build a new dataset DIALSTORY, which consists of 105k Chinese stories with a large amount of dialogue weaved into the plots to support the evaluation. We show the difficulty of the proposed tasks by testing existing models with automatic and manual evaluation on DIALSTORY. Furthermore, we propose to learn explicit character representations to improve performance on these tasks. Extensive experiments and case studies show that our approach can generate more coherent and informative dialogue, and achieve higher speaker recognition accuracy than strong baselines.

1 Introduction

Dialogue plays an important role in various types of literary works such as short stories, novels, and screenplays by advancing plots, establishing characters, and providing expositions using natural and lifelike words (Kennedy et al., 1983). Compared to dialogue in conversational scenarios such as chit-chat bots (Shang et al., 2015) or task-oriented dialogue systems (Deng et al., 2012), dialogue in stories is mainly used to exhibit emotions, motivations, or personalities of characters following the authors' design, which further serve for the coherence, informativeness, engagingness, and plot development of whole stories. Dialogue is also revealed to be essential for user-agent interaction in many text adventure games (Xi et al., 2021;Li et al., 2022). It has not been widely explored for machines to understand and generate dialogue in stories despite the broad recognition of the importance of this ability.



Context: Spring is here. The little squirrel, who had been in a hole in the tree for the winter, happily came out of the hole and jumped up and down among the trees. *（春天来了。在树洞中局促了一个冬天的小松鼠高兴地钻出洞来，在树林间翻腾跳跃起来。）*

Dialogue: "Come back, baby!" *（"孩子，回来！"）* **Speaker:** Mother squirrel

Context: The mother squirrel called out to the baby squirrel anxiously and said. *（松鼠妈妈着急地呼喊小松鼠说，...）*

Dialogue: "Spring has just begun, it is still very cold, be careful of freezing!" *（"刚开春，天还很冷，小心冻着！* **Speaker:** Mother squirrel

Dialogue: "I want to play a little longer." *（"我再玩一会儿！"）* **Speaker:** baby squirrel

Context: In spite of his mother's advice, the little squirrel swung on the branch again. Only after his mother yelled again did the squirrel start to walk home slowly. However, just before he reached the door, the squirrel sneezed and sneezed. 不顾妈妈的劝阻，小松鼠又在树枝上荡起了秋千。在妈妈的再三吆喝下，小松鼠才开始慢腾腾地往家里走。可是，刚走到门口，小松鼠就阿嚏阿嚏地打起了喷嚏。

Dialogue: "You caught a cold because you disobeyed me." *（"你看看，不听话，这不感冒了！"）* **Speaker:** mother squirrel

Context: The mother squirrel first let the little squirrel drink some water, and said: *（松鼠妈妈先让小松鼠喝了些水，然后说：）*

Dialogue: "[MASK]" *（"[MASK]"）* **Speaker:** mother squirrel

Context: When coming to an overgrown mound, the mother squirrel stomped three times on the mound. The mother rabbit heard and came out of the grass. *（来到一处杂草丛生的土坎下时，松鼠妈妈咚咚咚地在土坎上跺了三下脚。一会儿，听到信号的兔妈妈从草丛中钻出来了。）*

Dialogue: "What can I do for you" *（"找我有事吗？"）* **Speaker:** mother rabbit

Context: The mother rabbit asked with concern. Mother Squirrel said: *（兔妈妈关心地问，松鼠妈妈说：）*

Dialogue: "I remember when I was a kid, we played together. Once I caught a cold. You helped me find a little yellow flower and after I ate it, the cold was cured. What kind of flower is that, and where can I find it? *（"记得小时候，我们在一块玩，有一次我感冒了，是你帮我找来一种小黄花吃了后，感冒才好的。那种花是什么花来着，在哪里能找到呢？"）* **Speaker:** mother squirrel

Context: After the mother squirrel finished speaking, she told the mother rabbit about the little squirrel's cold. *（松鼠妈妈说完，把小松鼠的情况告诉了兔妈妈。）*

Dialogue: "That flower is called Forsythia, and there are many on the hillside! Let's go for it together!" *（"那种花叫连翘花，在半山坡上有很多呢！走，我们一块找去！"）* **Speaker:** [MASK]

Dialogue: "Thank you, no need! In broad daylight, you can't climb trees. Once you are chased by the hounds, you will be in trouble. I will do it myself!" *（"谢谢你，不用了！这大白天的，你不会爬树，一旦遇到猎狗追赶就麻烦了，我自己去就行！"）* **Speaker:** mother squirrel

Dialogue: "Stay at home and I'll get you some medicine." *（"在家里好好待着，我给你找点药去！"）*

Figure 1: An example from DIALSTORY and the proposed two tasks: *Masked Dialogue Generation* and *Dialogue Speaker Recognition*.

To spur research in this field, we present a new story dataset named DIALSTORY, which consists of 105k Chinese short stories with automatic annotations of dialogue turns and corresponding speakers. Furthermore, we formulate two new tasks including: *(1) Masked Dialogue Generation (DialGen):* completing a story with several dialogue turns masked with placeholders; and *(2) Dialogue Speaker Recognition (DialSpk):* choosing correct speakers from given candidates for several specified dialogue turns. We construct standardized datasets for these tasks through automatic or manual annotation based on DIALSTORY. As exemplified in Figure 1, these tasks comprehensively investigate the ability to capture characters' emotions (e.g., the *mother squirrel* is worried about baby squirrel's catching cold), motivations (e.g., the *mother squirrel* intends to call her baby to come

[*] Corresponding author

back home or get medicine), and relationship between each other (e.g. *mother rabbit* knows the kind of flower as a friend of *mother squirrel* when they were kids).

Furthermore, we found in massive stories that dialogue had a strong connection to different characters, reflected in their emotional states, speaking styles, and plot development. To provide more insights to tackle these tasks, we propose to learn character representations for modeling dependencies between characters and dialogue explicitly. Extensive experiments and case studies show that our approach can generate more coherent and informative dialogue, and achieve higher speaker recognition accuracy than strong baselines.

2 Related Works

Story Understanding and Generation Recently there have been various tasks proposed to assess the ability of story understanding and generation such as story ending selection (Zhou et al., 2019), story ending generation (Guan et al., 2019), story completion (Wang and Wan, 2019), story character identification (Brahman et al., 2021), story generation from prompts (Fan et al., 2018), titles (Yao et al., 2019) and beginnings (Guan et al., 2020). Another line of work focused on controllable attributes in story generation such as keywords (Xu et al., 2020), outlines (Rashkin et al., 2020), emotional trajectories (Brahman and Chaturvedi, 2020), styles (Kong et al., 2021) and characters' personalities (Zhang et al., 2022). Different from them, we emphasize the importance of dialogue in developing characters and maintaining the story's coherence.

Dialogue in story The dialogue context is comprised of a variety of messages from users and the system, determining the conversation topic and the user's goal for the conversation(Serban et al., 2017). The dialogue in the story makes this feature more apparent. Recently, like AI Dungeon, some text adventure games focus on players and machines co-creating stories. Li et al. (2022) proposed the idea that players can play the role of the chosen character and chat with others through the dialogue, while Xi et al. (2021) presented an AI game where players can explore different ways to reach the plot goals. In contrast, our tasks focus on both the understanding and generation of dialogue in the story, and they can be easily generalized to more complex AI interactive games.

Character Representations Dialogue unfolds in a story around several inter-related characters. Accordingly, it is essential to capture the traits of different characters for dialogue modeling. Ji et al. (2017); Clark et al. (2018) proposed to update characters' representations dynamically based on previous hidden outputs to track their states, which does not apply to the parallel architecture of Transformer for training. Azab et al. (2019) derived character representations from their corresponding dialogue turns. In contrast, we learn character representations from story plots, which then serve to understand and generate dialogue.

3 DIALSTORY Dataset

We construct the DIALSTORY dataset by randomly sampling 105k chapters from the Chinese novels released by Guan et al. (2022) with each chapter including at least ten dialogue turns. We also set a restriction that the number of tokens in all dialogue turns should account for at least 30% and at most 50% of the total length of the story, in order to keep a balance between the context and dialogue. We automatically annotate dialogue turns in these stories as text spans that are surrounded by quotation marks. Then, we use a pretrained named entity recognition model (Zhao et al., 2019) to identify all people's names. Each distinct name corresponds to a character. We also conduct a manual annotation on 150 stories, and the accuracy of character identification is 718/746=96.2%, which shows the high quality of this automatic method. We then decide the speaker of the dialogue by recognizing the subjects of sentences before and after the dialogue turn using spaCy[1]. Table 1 shows the statistics of our dataset.

DialStory	
Avg. #Token	451.98
Avg. #Dialogue Token	20.68
Avg. #Sentence	22.14
Avg. #Dialogue Turn	12.78
Avg. #Character	3.54

Table 1: Statistical average numbers for the DIALSTORY dataset. *#Dialogue token* means the average number of tokens in each dialogue turn.

[1] https://spacy.io/

Statistics	DialGen Training	 Validation	 Test	DialSpk Training	 Validation	 Test
#Examples	100k	2.5k	2.5k	20k	100	150
Avg. #Token (Input)	453.64	454.12	453.77	489.51	489.25	490.49
Avg. #Sentence (Input)	21.75	21.92	21.71	28.44	28.14	28.75
Avg. #Dialogue Turns (input)	10.35	10.45	10.34	13.68	13.54	13.82
Avg. #Token (Output)	82.04	81.94	81.64	-	-	-
Avg. #Dialogue Turns (Output)	3.91	3.90	3.94	-	-	-
Avg. #Specified Dialogue Turn	-	-	-	5.01	5.05	4.85
Avg. #Candidate Character	-	-	-	5.27	5.82	4.97

Table 2: Statistics for the DialGen and DialSpk dataset.

4 Proposed Tasks

We aim to measure model's ability to understand and generate dialogue in a story. To this end, we design the dialogue generation task Masked Dialogue Generation and dialogue understanding task Dialogue Speaker Recognition. We show the task definitions, targets, dataset construction and statistics below.

4.1 Masked Dialogue Generation

Task Formulation We formulate the DialGen task as follows: given an incomplete story $S = (s_1, s_2, \cdots, s_T)$ where each s_i is a token and several dialogue turns are masked with placeholder tokens [MASK], the model should generate the missing dialogue turns consecutively to form a coherent story. We denote the ground truth as $D = (d_1, d_2, \cdots, d_N)$ of N tokens.

Dataset Construction We use the following constraints to construct the DialGen dataset based on DIALSTORY:

- We randomly mask 30% of the dialogue turns in each story.
- We do not mask the first 50 tokens to provide sufficient background information for the story.
- We do not mask the last 30 tokens to provide ending information for that story.
- We ensure that each input story (i.e. with masked dialogue turns) mentions at least five characters.

Table 2 shows the detailed statistics.

4.2 Dialogue Speaker Recognition

Task Formulation We formulate the DialSpk task as follows: given a story $S = (s_1, s_2, \cdots, s_T)$ and a set of all character names $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$ that are mentioned in S as candidates, the model should choose the correct speakers from $\mathcal{C}$ for M specified dialogue turns $\mathcal{D} = \{D_1, D_2, \cdots, D_M\}$ in S.

Dataset Construction We randomly sampled 20k stories from DIALSTORY and automatically annotate the speaker for each dialogue turn for training, and resorted to manual annotation for validation and testing. For manual annotation, we first ask one annotator to label the characters in a story and the speaker of each dialogue turn. Then we asked another two annotators to check the correctness of the annotations, e.g., whether all mentioned characters are annotated, and whether each dialogue speaker is correct. We require the first annotator to re-annotate those examples that another two annotators do not agree on, and repeat the above process until all annotators agree on the examples. We also sampled 100 stories in the training set for manual annotation to investigate the accuracy of automatic annotation, which we will discuss in Section 6.2. Table 2 shows the detailed statistics.

5 Methodology

We propose to learn representations of different characters and exert them on decoding masked dialogue turns or predicting speakers. In this section, we describe the details of our model. Figure 2 shows the model overview for the DialGen task.

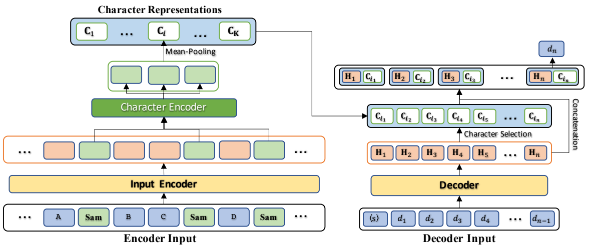

Character Representations
$\mathbf{C}_1$
...
$\mathbf{C}_i$
...
$\mathbf{C}_K$
Mean-Pooling
Character Encoder
...
...
Input Encoder
...
A
Sam
B
C
Sam
D
Sam
...
Encoder Input
d_n
$\mathbf{H}_1$ $\mathbf{C}_{i_1}$
$\mathbf{H}_2$ $\mathbf{C}_{i_2}$
$\mathbf{H}_3$ $\mathbf{C}_{i_3}$
...
$\mathbf{H}_n$ $\mathbf{C}_{i_n}$
$\mathbf{C}_{i_1}$ $\mathbf{C}_{i_2}$ $\mathbf{C}_{i_3}$ $\mathbf{C}_{i_4}$ $\mathbf{C}_{i_5}$ ... $\mathbf{C}_{i_n}$
Character Selection
Concatenation
$\mathbf{H}_1$ $\mathbf{H}_2$ $\mathbf{H}_3$ $\mathbf{H}_4$ $\mathbf{H}_5$ ... $\mathbf{H}_n$
Decoder
⟨s⟩ d_1 d_2 d_3 d_4 ... d_{n-1}
Decoder Input

Figure 2: Model overview for the DialGen task. ⟨s⟩ is the start-of-sequence token.

5.1 Character Representation Learning

We derive the representation of a character C_i, denoted as $\mathbf{C}_i$, by aggregating the output hidden states of the BART encoder corresponding to all mentions of C_i as follows:

$$\mathbf{C}_i = \text{Pool}(\text{CharEnc}(\{\mathbf{h}_k | s_k \text{ mentions } C_i\})), \quad (1)$$

where CharEnc is a bi-directional Transformer character encoder, Pool means applying mean pooling on the character encoder outputs, $\mathbf{h}_k$ is the hidden state of the input encoder at the k-th position with s_k as the input token. In that way, we get representations of all the characters from the input: $\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_K$, where K is the total number of the characters appearing in the text. And these representation vectors can be used at the decoding stage or other comprehension tasks.

5.2 Character Representation Utilization

DialGen We incorporate learned representations of K characters into the decoder to build explicit connections between dialogue and character features. At the n-th time step, the decoder dynamically selects a character C_{i_n} and then combines its representation $\mathbf{C}_{i_n}$ for token prediction as follows:

$$P(d_n | d_{<n}) = \text{softmax}(\boldsymbol{W} f([\mathbf{H}_n; \mathbf{C}_{i_n}]) + \boldsymbol{b}), \quad (2)$$

where $\mathbf{H}_n$ is the output hidden state of the decoder at the n-th step, $[;]$ is the concatenation operation, $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters and f is the SiLU activation function (Elfwing et al., 2018) followed by layer normalization. The decoder decides C_{i_n} as follows:

$$i_n = \text{argmax}_i \hat{\mathbf{H}}_n \cdot \mathbf{C}_i, \quad i = 1, 2, \cdots, K \quad (3)$$

$$\hat{\mathbf{H}}_n = \boldsymbol{W}_c \mathbf{H}_n + \boldsymbol{b}_c, \quad (4)$$

where $\mathbf{C}_i$ is the representation for C_i, $\boldsymbol{W}_c$ and $\boldsymbol{b}_c$ are trainable parameters. Finally, the model is optimized by minimizing the standard language modeling loss as follows:

$$\mathcal{L}_{\text{DialGen}} = -\sum_{n=1}^{N} \log P(d_n | d_{<n}). \quad (5)$$

DialSpk We feed the input story S into the input encoder and obtain the representations of all candidate characters following Eq. 1. We use the output hidden state at the position of a special token $[\text{MASK}]$, which is inserted before each specified dialogue turn D_m $(m = 1, 2, \cdots, M)$, as the corresponding dialogue representation $\mathbf{D}_m$. Then we optimize the following loss:

$$\mathcal{L}_{\text{DialSpk}} = -\sum_{m=1}^{M} \log P(\hat{C}_{i_m} = C_{i_m}), \quad (6)$$

$$P(\hat{C}_{i_m}) = \text{softmax}(\mathbf{D}_m \cdot \{\mathbf{C}_k\}_{k=1}^{K}), \quad (7)$$

where C_{i_m} is the ground-truth speaker label for D_m $(i_m = 1, 2, \cdots, K)$, $P(\hat{C}_{i_m})$ is the distribution of the predicted speaker overall the candidate list $\mathcal{C}$.

6 Experiments

Since our approach adapts to all encoder-decoder models, we build our model based on $\text{BART}_{\text{Base}}$ (Shao et al., 2021). We use $\text{BART}_{\text{Base}}$, BERT(Devlin et al., 2019), RoBERTa(Liu et al., 2019), and MacBERT(Cui et al., 2020) as baselines. We follow $\text{BART}_{\text{Base}}$'s hyper-parameters and initialize our model with the public checkpoint[2]. Both the input encoder and decoder contain 6 hidden layers with 768-dimensional hidden states. BERT[3], RoBERTa[4], and MacBERT[5] uses 12-layer encoder with 768-dimensional hidden states. All the models are trained on Quadro RTX 6000 GPU.

6.1 Masked Dialogue Generation

Implementation Details To conduct experiments on the masked dialogue generation task, we decide the hyper-parameters based on the performance of the validation set. We train Shao et al. (2021)'s BART model for 4.6 epochs with a 1e-4 learning rate for 1 day, and for our model, we train

[2] https://huggingface.co/fnlp/bart-base-chinese

[3] https://huggingface.co/bert-base-chinese

[4] https://huggingface.co/hfl/chinese-roberta-wwm-ext

[5] https://huggingface.co/hfl/chinese-macbert-base

Model	#Param	BLEU1	BLEU2	DIST2	DIST3	DIST4
BART	116M	21.04	7.85	6.67	20.57	33.78
Ours	125M	**23.88**	**8.47**	**7.72**	**25.35**	**43.10**

Table 3: Automatic evaluation results for DialGen. #Param is the number of parameters.

Models	Fluency(%)				Coherence(%)				Informativeness(%)			
	Win	Lose	Tie	κ	Win	Lose	Tie	κ	Win	Lose	Tie	κ
Ours vs. BART	30.0	30.0	40.0	42.0	39.9**	15.0	45.1	42.3	41.0*	17.0	42.0	43.5

Table 4: Manual evaluation results. The scores indicates the percentage of *win, lose*, or *tie* when comparing our model with the BART baseline. κ denotes Fleiss' Kappa (Fleiss, 1971) to measure the inter-annotator agreement. * means p-value < 0.05, ** means p-value < 0.01 (Wilcoxon signed-rank test).

it for 5.6 epochs with a 1e-4 learning rate for 1 day. All baselines and our model are trained using the Adam optimizer.

During the training process for our method, we computed the selection coverage of characters within a single story. And it showed that in every 1000 training steps, the coverage of different characters ranged from 98.64% to 99.00%, which meant nearly all the characters are selected during training, and all the characters contributed to the generated dialogue. It further proved that the argmax in Eq. 3 operation doesn't break the gradient progress when training for this task.

Automatic Evaluation Following previous works, we use several standard, widely used automatic evaluation metrics. We use **BLEU-**n (Papineni et al., 2002) to measure the average word overlap between each generated and ground-truth dialogue turn (n=1,2), and **Distinct-**n (Li et al., 2015) to evaluate n-gram diversity of generated dialogue turns (n=2,3,4).

We further train a classifier to evaluate the overall **coherence** by fine-tuning BERT$_{\text{Base}}$ (Devlin et al., 2018) to discriminate stories whose dialogue turns are randomly reordered (labelled with 0) from original stories (labelled with 1) (Guan and Huang, 2020).

To be more specific, for the coherence classifier, we construct the training and validation sets by randomly shuffling the order of dialogue turns and keeping other content in the correct order. We regard the perturbed story as a negative example and the original story as a positive example. We sample another 195k stories (except those in DIALSTORY) from the novels of Guan et al. (2022) to construct the training set (190k examples) and the validation set (5k examples). We train the model for 4 epochs with a 2e-5 learning rate and a 16-batch size, using the Adam optimizer. During the evaluation, we consider an example coherent when the probability of being coherent predicted by the classifier is greater than 0.5. We use the ratio of outputs (along with the input) that are classified as coherent by the classifier to all generated outputs as the coherence score.

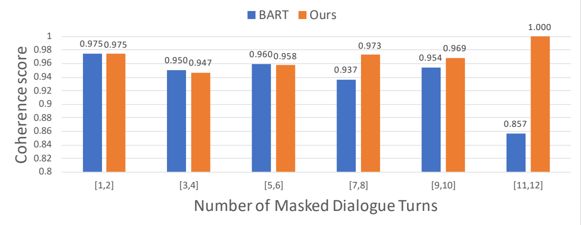


Figure 3: The coherence score of our model and BART varies with the number of masked dialogue turns.

The result of the automatic evaluation is presented in Table 3. According to the table, compared to the BART baseline, our model consistently generates more word overlaps with ground truth and achieves better diversity under the guidance of character representations, which means our model can generate more diverse but not commonplace responses.

Figure 3 plots the coherence score varying with the number of masked dialogue turns. The result shows that our model gets a higher coherence score than BART when required to generate more than seven turns of dialogue in one story.

Manual Evaluation We conduct a pairwise comparison between our model and the BART baseline. We randomly select 100 examples from the test set. For each pair of outputs along with the input, we ask three annotators to give a preference (*win*, *lose* and *tie*) in terms of fluency, coherence, and informativeness. All the annotations are native Chinese speakers. We adopt majority voting to make final decisions among the annotators. The three aspects of manual evaluation are as follows:

1. **Fluency:** Grammatical correctness and intra-sentence linguistic quality.

2. **Coherence:** Inter-sentence relatedness, causal and temporal dependencies. We judge the coherence between the story and a dialogue turn by following the criterion in Table 5. We add the scores of all the generated dialogue turns in a story to get the overall coherence score of the story, which is then used to compare with each other.

3. **Informativeness:** Interesting, diverse and rich details.

Score	**Coherence Criterion**
0	has nothing to do with the whole story
1	with correct mentions or reference but not logical
2	perfectly match the masked position

Table 5: Coherence scoring criterion for each generated dialogue turn.

As shown in Table 4, all the results show moderate ($\kappa > 0.4$) agreement, which shows our model outperforms the BART baseline significantly in dialogue informativeness and coherence.

Case Study Figure 5 showed two examples to investigate how learning character representations can help our model generate more coherent dialogue. We found that our model can better model the relationship between different characters and the direction of the storyline. For example, in the first case, we can see that the BART's generation confuses different characters' fathers, while our model captures the relationship between different characters, and generates proper responses for the corresponding characters, which also moves the plot forward. And in the second case, we can see that BART's generation is commonplace and contradicts the plot development. In contrast, our model captures the intentions of the speaker and the development trend of the plot, generating an appropriate and coherent response. Since these two models use the same pretrained weight, we can infer that the character modeling module leverages the coherent and reasonable generation.

We also summarize four error types of the generated dialogue turn for the DialGen task: **(1)** Inter-sentence Contradiction; **(2)** Inter-sentence Repetition; **(3)** Intra-sentence Contradiction; **(4)** Intra-sentence Repetition. We show the typical corresponding cases in Figure 4. We conducted a quantitative analysis of those 4 error types on our model's generation. We analyzed 20 stories with 103 dialog turns and the results are shown in Figure 6. We found that both our model and BART suffer from these errors, suggesting that there is still space for model improvement, especially in the inter-sentence repetition.

Error Type 1	**Inter-sentence Contradiction**
Story	**Ziqing** looked around, looking for **Yuehua**. **Jianguo** told him, "Stop looking for her!". **Ziqing** replied, **"[MASK]"** 子晴看看周围寻找月华，建国告诉他：“别再找月华了”子晴回答到**“[MASK]”**
Output	I have found **Yuehua**. 我找到月华了。
Error Type 2	**Inter-sentence Repetition**
Story	**Zhuge** said,"It's just a little complicated. The real treasure thief is about to appear." And he then whispered: **"[MASK]"** 诸葛明道“只是有些牵连。真正的盗宝人，就快出现了。”一面又低声道：**“[MASK]”**
Output	The real treasure thief is about to appear. 真正的盗宝人要出现了
Error Type 3	**Intra-sentence Contradiction**
Output	We are not unrighteous people, we are unrighteous people. 我们不是无耻之人，是无义之人。
Error Type 4	**Intra-sentence Repetition**
Output	What use is it? What use is it? What use is it? 这有什么用？这有什么用？这有什么用？

Figure 4: Cases for different error types for the DialGen task.

Error Type	Type 1	Type 2	Type 3	Type 4
Proportion	4.85%	16.5%	1.94%	2.91%

Table 6: The quantitative analysis of the error types.

6.2 Dialogue Speaker Recognition

Implementation Details To conduct experiments on the speaker recognition task, we decide the hyper-parameters based on the performance of the validation set. For the BART baseline and our approach, we insert a mask token before each dialogue needed to be predicted, and a person id token

Story	**Xu** immediately said in a hurry, “In this case, please go back, so that the old gang leader and the old lady don’t worry about you!” **Eliza** didn’t want to go back, and immediately said disapprovingly, “It’s okay, it must be my father who met these two stubborn old men...” **Xu** still insisted that “**[MASK1]**”. **Marry** was afraid that the old gang leader would scold her for not persuading **Eliza** to go back earlier, so she said sternly, “Girl, let’s go back quickly. **Xu** will come back after the business!” **Eliza** heard it, and had no choice but to say “**[MASK2]**”. **Xu** was relieved when she heard it, and immediately smiled, “See you around!” 许先生赶紧说：“这种情况先回去，这样头目和老太太就不会担心你！”艾丽莎不想回去，立刻失望地说道：“这也可以，但见了这两个人的我父亲......”许先生仍然说道**“[MASK1]”**玛丽害怕头目会责怪她没有早点说服艾丽莎回去，所以她坚决地说道：“女孩，让我们赶快回去。许先生会在生意结束之后回来。”艾丽莎听了这话，没有其它选择只能说：**“[MASK2]”**当她听了这话，许先生终于放心了，立刻笑着说：“待会儿见！”
BART	**[MASK1]**: No, **my father and old lady** are not two old stubborn. 不，我父亲和老太太两个不是顽固的人。 **[MASK2]**: Girls, let’s go back, the old gang leader **has already told us!** 女孩，让我们回去，老头目已经告诉过你了！
Ours	**[MASK1]**: The old lady won’t blame you, **as long as you go back**, nothing else will happen! 老太太不会责怪你的，只要你回去就当是无事发生！ **[MASK2]**: Okay, **let’s go back first**! 好的，让我们先回去！
Story	**The conductor** was not happy, but rather melancholy. He stood for a while, then said, “I think he must be punished now. I’m going to fire him.” “Fire him? What’s the benefit of that? Do you think it would be wiser to teach him how to treat passengers better and keep him in the future?” **“[MASK1]”**, “He insulted the old in front of all these people. Shouldn’t he be called in and apologize in front of everyone?” “OK, I’ll call him here.” 售票员不太开心，而是忧郁。他站了一会儿，然后说：“我认为他必须立刻受到惩罚。我打算开除他。” “开除他？那有什么好处？你认为会有比留下他并教他怎么更好地对待乘客更明智的做法吗？” **“[MASK1]”** “他在这些人面前侮辱了那个老人。难道他不应该被叫来并在所有人面前道歉吗？” “好的，我叫他过来。”
BART	**[MASK1]**: I don’t think so. 我不这样认为。
Ours	**[MASK1]**: I think he should be punished now. 我认为他现在需要被惩罚。

Figure 5: Case study of the DialGen task. The special tokens [MASK1], [MASK2] refer to the positions of missing dialogue turns. We highlight the different characters in bold.

before and after each character name span. Then, we insert all the unique person id tokens before the input stories as different options, and make predictions based on the cos similarity of option tokens and mask tokens. We train Shao et al. (2021)’s BART model for 30 epochs with a 5e-5 learning rate for 3 days. For encoder-only baselines, we implemented BERT, RoBERTa, and MacBERT and trained them for 15 epochs with a 1e-5 learning rate for 2 days. For our model, we train it for 22 epochs with a 1e-6 learning rate for 2 days. All baselines and our model are trained using the Adam optimizer.

Metrics We evaluate the DialSpk task using two automatic metrics including dialogue-level accuracy (DAC) and story-level accuracy (SAC). DAC is calculated as the ratio of the correct predictions to the total number of specified dialogue turns, while SAC is the ratio of the number of stories where all dialogue turns are correctly predicted to the number of all test examples. These two metrics provide the evaluation for dialogue understanding with different granularities.

$$\textbf{DAC} = \frac{\text{\# correct predictions}}{\text{\# total number of specific dialogue turns}} \times 100\% \tag{8}$$

$$\textbf{SAC} = \frac{\text{\# correct stories}}{\text{\# total number of stories}} \times 100\% \tag{9}$$

Results As shown in Table 7, our model outperforms all the baselines significantly ($p < 0.01$, Wilcoxon signed-rank test) on both DAC and SAC scores, suggesting the benefit of learning character representations. We tested the accuracy of automatic training set annotations, and the DAC/SAC scores are 86.78%/67.80%. Together with the model’s performance on the test set, we can see

Models	#Param	DAC(%)	SAC(%)
BERT	103M	62.00	20.00
RoBERTa	103M	61.80	21.30
MacBERT	103M	61.70	20.00
BART	116M	61.80	21.20
Ours	125M	**93.30**	**74.70**

Table 7: Experiment results for the DialSpk task.

the automatic annotation for the training set is of good quality. We also conducted the human prediction experiment, and the DAC/SAC scores are 97.90%/90.70%, which are much higher than the best model. So there is much room for further improvement for machine-based approaches.

7 Discussion

Masked Dialogue Generation In this task, the masked positions can be anywhere in the story. To generate and complete one specific masked dialogue turn, this task requires the machine to first understand the main line of the whole story, the roles and features of different characters, and then infer and generate the most appropriate dialogue turn to move the plot forward according to the specific environment. Context information also plays an important role in this task, which puts forth a high demand for dialogue and plot coherence. And this task can easily leverage other research such as persona chat bots, emotional chat bots, or even some other story generation tasks.

Dialogue Speaker Recognition This task specifically focuses on the dialogue understanding in a story. In a story, there are always lots of characters speaking to each other in different environments. And the relationship between characters and dialogue is complicated. Sometimes a piece of dialogue is self-talk, sometimes it is a conversation between two parties, or even multiple parties in a complex environment, and sometimes there are many details in the background information that cover up the actual speaker. The speaker of the dialogue will not only appear in the surrounding, but may also hide in the previous or behind the context of the dialogue. To recognize all the speakers in one story, our model also needs to capture different characteristics of people. Higher requirements for the model's understanding of dialogue and character relationships are also put forth by the necessity to predict speakers simultaneously in several positions. This task can also leverage other research on recognition in multi-person dialogue story scenes.

8 Future Work

Despite the overall improved performance of our character modeling methods, we find that there is still space for further improvement. Our approach constructs the character representations from the input stories as static vectors. But the status of characters could be updated during the generation process. As a result, we can dynamically update those representations during plot development. We plan to explore the new possibility of the character-update technique and leave it as an important future work.

9 Conclusion

In this work, we present the first study on understanding and generating inter-character dialogue in stories. To this end, we collect a Chinese story dataset DIALSTORY with a large amount of dialogue, and propose two new tasks including masked dialogue generation and dialogue speaker recognition. We also construct standardized datasets for these tasks through automatic and manual annotations based on DIALSTORY. By incorporating representations of different characters, our model outperforms strong baselines significantly on both tasks in terms of automatic and manual evaluation. The benchmark datasets, tasks, and models will further boost the development of this field.

10 Limitations

For the DialSpk task, in the test set, there are 150 stories, and a total of 728 masked dialogue positions, and in the validation, there are 100 stories and 505 positions. For the DAC score, there is enough data to evaluate the models' performance. The dataset is small to some degree for the SAC evaluation. Although the validity of our model could be reflected in this dataset, we also plan to augment this annotated dataset for future research.

References

Mahmoud Azab, Noriyuki Kojima, Jia Deng, and Rada Mihalcea. 2019. Representing movie characters in dialogues. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 99–109, Hong Kong, China. Association for Computational Linguistics.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. “let your characters tell their story”: A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elizabeth Clark, Yangfeng Ji, and Noah A Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pretrained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Li Deng, Gokhan Tur, Xiaodong He, and Dilek Hakkani-Tur. 2012. Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 210–215. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. Lot: A story-centric benchmark for evaluating chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, 10:434–451.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Jian Guan and Minlie Huang. 2020. UNION: an unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9157–9166. Association for Computational Linguistics.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A Smith. 2017. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839.

XJ Kennedy, Dana Gioia, and Dan Stone. 1983. *Literature: An introduction to fiction, poetry, drama, and writing*.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. Stylized story generation with style-guided planning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Wanshui Li, Yifan Bai, Jiaxuan Lu, and Kexin Yi. 2022. Immersive text game and personality classification. *arXiv preprint arXiv:2203.10621*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Tianming Wang and Xiaojun Wan. 2019. T-CVAE: transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5233–5239. ijcai.org.

Yadong Xi, Xiaoxi Mao, Le Li, Lei Lin, Yanjiang Chen, Shuhan Yang, Xuhan Chen, Kailun Tao, Zhi Li, Gongzheng Li, et al. 2021. Kuileixi: a chinese open-ended text adventure game. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 175–184.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2831–2845. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. Persona-guided planning for controlling the protagonist's persona in story generation. *arXiv preprint arXiv:2204.10703*.

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. *EMNLP-IJCNLP 2019*, page 241.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2019. Story ending selection by finding hints from pairwise candidate endings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):719–729.