Generative Chatbot Framework for Cybergrooming Prevention

Pei Wang

Master Thesis Defense Exam

Committee Members:
Jin-Hee Cho, Chair
Lifu Huang, Co-chair
Chang-Tien Lu, Member

Dec. 1, 2021
Outline

- Motivation & Goal
- Key Contributions
- Related Work
- The Proposed Generative Chatbot Framework: SERI
  - Cybergrooming Stage Classification
  - Chatbot Pre-training
  - Chatbot Fine-tuning
- Experiment Setup
- Experimental Results & Analysis
- Conclusions
Motivation & Goal

- **Cybergrooming**: Luring people, particularly children or young adults, for sexual exploitation in cyberspace.

- The majority of cybergrooming studies have focused on detecting predators.

- However, there has been a lack of studies to proactively protect potential youth victims from cybergrooming.

- **Goal**: Develop a generative chatbot framework that can provide authentic conversations between a perpetrator chatbot and a youth user.
Key Contributions

- **Applied a two-stage paradigm:**
  - **Pre-training** on general and large-scale causal talk datasets
  - **Fine-tuning** on the target dataset

- Employed **deep reinforcement learning** (DRL) to augment the quality of the dialogue model that can generate strategic conversation with cybergrooming goals.

- Developed a mechanism to **escalate the attack stages** and coordinated the dialogue generation.

- Evaluated the chatbots by using both referenced metrics and unreferenced metrics including **human evaluation**.
Related Work

- **Detection of perpetrators’ languages** by leveraging various machine learning algorithms:
  - $k$-nearest neighbors (KNN) (Gunawan et al., 2018)
  - Support Vector Machine (Anderson et al., 2019)
  - Naïve Bayes (Bours and Kulsrud, 2019)
  - Neural Network (NN) classifiers (Fauzi and Bours, 2020)

- **Chatbot applications**:
  - SnapTravel (Hooijdonk et al., 2019)
  - Google Hangouts (Google, 2021)
  - Guardian (Good et al., 2021)

- **Chatbot applications for cybergrooming prevention**:
  - Negobot (Laorden et al., 2013)
Related Work

- **Pre-training language models:**
  - RNN (B. Quast, 2016)
  - Transformer (T. Wolf et al., 2020)
  - GPT (Radford et al., 2018, 2019)
  - BERT (Devlin et al., 2019)
  - T5 (Raffel et al., 2020)

- **DRL-based conversation generation:**
  - Markov decision process (MDP)
  - Reinforcement Learning (RL)
  - Deep Reinforcement Learning (DRL)
Limitations

- Prior effort has focused on the grooming prevention without characterizing the features of victims by cybergrooming.
- Most chatbots have been developed on the top of RNN and GPT model, not T5.
- No previous research has developed a chatbot to generate conversations between a cybergroomer and a potential victim.
- No prior chatbots have considered goal-driven conversations to address perpetrators’ adversarial goal.
SERI: Architecture Overview

Perpetrator Chatbot

- **Step I**: TextCNN Classifier
- **Step II**: T5 Pre-train
- **Step III**: T5 Fine-tune

Victim Chatbot

- **Step I**: T5 Pre-train
- **Step II**: T5 Fine-tune

Architecture of the proposed SERI framework

**Step IV: Stage Evolution**

- **Stage $s_1$**: Greetings and casual talk
- **Stage $s_2$**: Collecting personal information
- **Stage $s_3$**: Sex-related contents
- **Stage $s_4$**: Attempting a personal encounter
### SERI: Cybergrooming Stage Classification

<table>
<thead>
<tr>
<th>Stages</th>
<th>Conversation Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\tilde{s}_1)</td>
<td>Greetings and casual talks to establish a trust relationship</td>
</tr>
<tr>
<td>(\tilde{s}_2)</td>
<td>Collecting private information, such as name, age, gender, location, interests, family, school, or schedule</td>
</tr>
<tr>
<td>(\tilde{s}_3)</td>
<td>Asking sexual questions or requests, talking about sexual conversations, or sending sexual pictures/videos</td>
</tr>
<tr>
<td>(\tilde{s}_4)</td>
<td>Attempting a personal contact or asking meeting in person</td>
</tr>
</tbody>
</table>

- Zambrano et al. (2019) labeled the cybergrooming dataset with the six stages. However, it was limited due to unclear distinctiveness between stages as many utterances from the perpetrator could fit multiple stages.
- We refine the six stages into a new set with four stages.
- We leverage the TextCNN model to train a stage classifier for the perpetrators.
Since the in-domain Perverted Justice (PJ) dataset is small, we improve the fluency of the generated conversations by pre-training T5 on the large-scale ConvAI2 dataset.

To train the T5-based chatbots, we take the following procedures:

- Concatenate two dialogue turns (i.e., 4 or 5 sentences) as a unit;
- Take the last sentence as the target one (i.e., ground truth response); and
- Treat the preceding sentences as the sources (i.e., dialogue history).

A sample training unit for the perpetrator and pseudo-user (i.e., potential victim) chatbots.
SERI: Pre-training on the ConvAI2 Dataset

- Given an input sequence $x$ as the source, a response is generated by optimizing the following objective:

$$\mathcal{L} = - \sum_i \log P(y_i|y_{i-k}, \ldots, y_{i-1}; x; \Theta),$$

where $\Theta$ denotes the set of parameters in the T5, and $y_i$ is the $i$-th token of the target response.

- We pre-train the perpetrator and the potential victim chatbots separately on the ConvAI2 dataset.

- We observe that the perpetrator chatbot tends to generate more leading dialogues while the potential victim chatbot generates response messages more consistently.
Dataset Segmentation: To obtain the messages for each stage, we cut conversations in the PJ dataset into several blocks and assign a stage for each block based on the criteria below.

<table>
<thead>
<tr>
<th>Stages</th>
<th>Label Distribution of Each Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{s}_1$</td>
<td>More than 80% utterances are labeled as $\tilde{s}_1$</td>
</tr>
<tr>
<td>$\tilde{s}_2$</td>
<td>More than 60% utterances are labeled as $\tilde{s}_2$</td>
</tr>
<tr>
<td>$\tilde{s}_3$</td>
<td>More than 50% utterances are labeled as $\tilde{s}_3$</td>
</tr>
<tr>
<td>$\tilde{s}_4$</td>
<td>More than 40% utterances are labeled as $\tilde{s}_4$</td>
</tr>
</tbody>
</table>

Conversation segmentation criteria for the four relationship stages.
SERI: Fine-tuning the Chatbots on the PJ Dataset - Rule-based Method: Subchatbot Fine-tuning

- **Chatbot fine-tuning:** We fine-tune the following:
  - The four perpetrator sub-chatbots on the four groups of blocks separately; and
  - A victim chatbot based on the victim utterances from the PJ dataset.

- **Drawbacks:**
  - Dataset segmentation will cut the whole training dataset into fragments; and
  - The threshold number is heuristic.
SERI: Fine-tuning the Chatbots on the PJ Dataset - DRL Policy Method: State, Action, and Reward

- **State:** A state is denoted by the two previous dialogue turns to contain four consecutive utterances $[u_1, u_2, u_3, u_4]$.
- **Action:** An action is a dialogue utterance to generate.
- **Reward:** We implement a classification confidence based reward to encourage the chatbot to follow the expected grooming states. The confidence of the stage classifier is estimated by:

$$p(s|y') = \text{softmax}(\text{TextCNN}(y', \theta)),$$
SERI: Fine-tuning the Chatbots on the PJ Dataset - DRL Policy Method: Calculation of Loss

Calculation of loss after integrating DRL.
SERI: Fine-tuning the Chatbots on the PJ Dataset - DRL Policy Method: Policy Gradients and Objectives

- **Gradients and objectives:** The reward is used for learning a policy. The policy gradient is given by:

\[
\nabla_{\Theta} J(\Theta) = E[R \cdot \nabla_{\Theta} \log P(y^s|x; \Theta)],
\]

where \( R = [p(s_i|y')] \),

- \( R \) is the stage classifier reward, the probability that the generated target sentence belongs to the correct stage from the ground truth. The \( y^s \) is sampled from the distribution of the model outputs at every decoding time step, and \( \Theta \) refers to the parameters of the model.
### Output filtering:
- Generate 5 candidate messages and select the best one.

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>what will we do if you meet me ?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Candidate messages</strong></td>
<td>1. meet you?, you want me to get in the car with you?, are you going to go?, my i’ll show you how to get it outta the car</td>
<td><strong>7.01</strong></td>
</tr>
<tr>
<td></td>
<td>2. i just got in my little bmw</td>
<td><strong>3.18</strong></td>
</tr>
<tr>
<td></td>
<td>3. well we will do it, that would be nice, that would be fun, i have to leave the car here and then i’ll go back my way</td>
<td><strong>7.61</strong></td>
</tr>
<tr>
<td></td>
<td>4. anything you want to do?, you want me to talk about it, and i want you to think of something</td>
<td><strong>8.95</strong></td>
</tr>
<tr>
<td></td>
<td>5. i won’t just go the store i got, i just wanna get everything outta the car so that it not be torrious</td>
<td><strong>5.98</strong></td>
</tr>
</tbody>
</table>
SERI: Stage Evolution of the Perpetrator Chatbot

- We observe whether each stage conversation maintains a certain number of rounds (e.g., 20).
- If the conversation of stage $\tilde{s}_1$ lasts 20 rounds, the perpetrator will move to stage $\tilde{s}_2$ with a trigger utterance.
- Once the victim detects grooming, he/she will leave the chat.
**SERI: Stage Evolution of the Perpetrator Subchatbot**

<table>
<thead>
<tr>
<th>Stages</th>
<th>Trigger Sentence of Each Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{s}_1$</td>
<td>hi, how are you doing today?</td>
</tr>
<tr>
<td>$\tilde{s}_2$</td>
<td>you parents know you be chatting with me?</td>
</tr>
<tr>
<td>$\tilde{s}_3$</td>
<td>how many pictures you have, any sexy?</td>
</tr>
<tr>
<td>$\tilde{s}_4$</td>
<td>what will we do if you meet me?</td>
</tr>
</tbody>
</table>

**Trigger sentences of the four relationship stages.**

- A trigger sentence can direct the conversation into the topic of a specific stage.
- **Demonstration** for the conversations of the chatbots
Experiment Setup - Datasets

- **The ConvAI2 dataset** (Dinan et al., 2019)
  2,000+ dialogues, 60K+ chat records.

- **The Perverted Justice (PJ) dataset** (PJ Website, 2010)
  100 dialogues, 100K+ chat records
Experiment Setup - Metrics

- **Referenced metrics**
  - BLEU (Post, 2018)
  - ROUGE-L (Lin, 2004)
  - BERTScore (Zhang et al., 2020)

- **Unreferenced metrics**
  - MaUde scores (Sinha et al., 2020)
  - Perplexity (Peter Brown et al., 1992)

- **Human evaluation**
  - Turing test (Turing, 2009)
**Referenced Metrics - BLEU**

- **BLEU** (Post, 2018): BLEU (Bi-Lingual Evaluation Understudy) score is used as one of fluency metrics and measures precision of different n-grams between target and references. Lower is more desirable.

\[
BLEU = BP \cdot \exp \left( \sum_{n=1}^{N} w_n \log P_n \right),
\]

where \( n \) denotes \( n \)-gram where \( N \) is the max \( n \)-gram order (default 4), \( w_n \) is the weight for different \( n \)-gram, \( P_n \) is the precision of the generation for different \( n \)-gram, and \( BP \) denotes Brevity Penalty. \( BP \) is given by:

\[
BP = \begin{cases} 
1 & \text{if } c > r \\
\exp^{1-r/c} & \text{if } c \leq r 
\end{cases}
\]

where \( c \) is the length of the target sentence, \( r \) is the length of the reference sentence which has the closest length to the target sentence.
Referenced Metrics - ROUGE-L

- **ROUGE-L** (Lin, 2004): ROUGE (Recall-Oriented Understudy for Gisting Evaluation)-L score also measures fluency of languages used and is calculated based on the longest common subsequence between the target and reference. Higher (close to 1) is more desirable.

\[
R_{lcs} = \frac{LCS(X, Y)}{m},
\]
\[
P_{lcs} = \frac{LCS(X, Y)}{n},
\]
\[
F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2P_{lcs}},
\]

where \(m\) is the length of reference sentence, \(n\) is the length of target sentence, and \(\beta\) is the weight coefficient for balancing recall and precision. In our calculate, we use the F-score with \(\beta\) equals to 1.
BERTScore (Zhang et al., 2020): BERTScore is a metric based on the pre-trained BERT model, computing BERT embeddings and pairwise cosine similarity between generated sentence and reference. For a candidate $\hat{x}$ and a reference $x$, the recall, precision and $F$-scores are:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j$$

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Higher is more desirable.
Referenced Metrics - MaUde and Perplexity

- **MaUde**: The MaUde score makes evaluation on the latent representations of utterances and indicates the extent of the reasonableness of the dialogues. MaUde uses large pre-trained language models to extract latent representations of utterances, and leverages the temporal transitions existing between them. Higher is more desirable.

- **Perplexity**: The perplexity score is an indicator of how to easily understand a given sentence. Perplexity as the normalized inverse probability of the test set is given by:

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Lower is more desirable.
Human Evaluation

The 200 conversation samples randomly selected evaluated by three human graders. Given 4 history utterances, the grader is asked to select a better response between 2 target utterances (i.e., one from the PJ dataset and the other generated by the SERI).
Referenced Metrics-based Analysis

<table>
<thead>
<tr>
<th>Role</th>
<th>BLEU</th>
<th>ROUGE</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max:100</td>
<td>Max:1</td>
<td>Max:1</td>
<td></td>
</tr>
<tr>
<td>Perpetrator</td>
<td>2.9906</td>
<td>0.0970</td>
<td>0.8311</td>
</tr>
<tr>
<td>Victim</td>
<td>2.6884</td>
<td>0.1063</td>
<td>0.8274</td>
</tr>
</tbody>
</table>

BLEU, ROUGE, and BERTScore-based analysis for the conversations generated by the SERI.

- BLEU and ROUGE scores reflect a low similarity between the generated and ground truth dialogues due to the distinctiveness of dialogue generation task.
- The BERTScore is relatively high because:
  - Most of the words are functional and uninformative, making the BERT difficult to learn meaningful contextual representations; and
  - The BERTScore is highly sensitive to some particular word pairs which do not capture any meaningful semantics of very short messages.
Unreferenced Metrics: MaUde Score-based Analysis

<table>
<thead>
<tr>
<th></th>
<th>Perpetrator</th>
<th>Victim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth dialogues</td>
<td>0.8442</td>
<td>0.8625</td>
</tr>
<tr>
<td>Generated dialogues</td>
<td>0.8662</td>
<td>0.8641</td>
</tr>
</tbody>
</table>

MaUde score-based analysis.

- Higher MaUde scores is more desirable.
- Higher MaUde scores observed in our generated dialogues imply that our chatbots can effectively mitigate the adverse effects of PJ dataset.
Unreferenced Metrics: Perplexity Score-based Analysis

<table>
<thead>
<tr>
<th></th>
<th>Perpetrator</th>
<th>Victim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth dialogues</td>
<td>357.06</td>
<td>477.82</td>
</tr>
<tr>
<td>Generated dialogues</td>
<td>139.46</td>
<td>188.97</td>
</tr>
</tbody>
</table>

Perplexity score-based analysis.

- Lower perplexity is more desirable.
- Lower perplexity scores are observed on our generated dialogues.
## Human Evaluation Analysis

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: nutting, you miss me</td>
<td></td>
</tr>
<tr>
<td>2: ya</td>
<td></td>
</tr>
<tr>
<td>3: you better</td>
<td></td>
</tr>
<tr>
<td>4: what if i don’t?, lol, jk</td>
<td></td>
</tr>
<tr>
<td>5: i’ll get you</td>
<td></td>
</tr>
<tr>
<td>6: can’t get me through the competition duh, i’m not scared of you</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Original response</th>
<th>Generated response</th>
</tr>
</thead>
<tbody>
<tr>
<td>lol, how much you miss me</td>
<td>i’m scared of you right now</td>
</tr>
</tbody>
</table>

Inter-agreement sample of human evaluation.

- The utterances generated by the SERI are chosen over human-written utterances by at least two annotators for 74 out of 200 samples, reaching a 37% passing rate for this Turing test (Turing, 2009).
### Pre-training Influence Analysis

<table>
<thead>
<tr>
<th>Metric</th>
<th>With Pre-training</th>
<th>W/O Pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>2.556</td>
<td>2.505</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.091</td>
<td>0.081</td>
</tr>
<tr>
<td>BERTScore</td>
<td>0.830</td>
<td>0.829</td>
</tr>
<tr>
<td>Perplexity</td>
<td>124.82</td>
<td>140.33</td>
</tr>
<tr>
<td>MaUde</td>
<td>0.853</td>
<td>0.850</td>
</tr>
</tbody>
</table>

Impact of pre-training on the ConvAI2 dataset.

- The results with all the five metrics indicate that the **model with pre-training** generates better dialogues than **model without pre-training**. Therefore, pre-training appears a helpful strategy to effectively train a chatbot.
Impact Analysis of Using DRL

<table>
<thead>
<tr>
<th></th>
<th>With DRL</th>
<th>W/O DRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>2.556</td>
<td>2.472</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.091</td>
<td>0.084</td>
</tr>
<tr>
<td>BERTScore</td>
<td>0.830</td>
<td>0.829</td>
</tr>
<tr>
<td>Perplexity</td>
<td>124.82</td>
<td>118.93</td>
</tr>
<tr>
<td>MaUde</td>
<td>0.853</td>
<td>0.8333</td>
</tr>
</tbody>
</table>

Impact Analysis Under the Models With or Without DRL.

- All five metrics except **Perplexity** indicate the model with DRL generate better dialogues than the model without DRL.
- Overall, integrating DRL introduces positive impacts to our chatbot.
Conclusions

1. Pre-training on a high-quality corpus first and then fine-tuning on the target corpus is a helpful strategy.

2. Training the dialogue model with accurate context utterances and target utterance can help distinguish the different roles of chatbots.

3. A grooming stage-based deep reinforcement learning method is helpful in evolving stages from the perpetrator perspective.

4. Our human evaluation also shows the promising performance of the SERI chatbots by reaching a 37% passing rate for the Turing test.
Limitations

1. It is still difficult to effectively reflect the logical fluency due to the inherent noises in the languages used in the context of cybergrooming.

2. Since cybergrooming lies in a unique situation in online chatting contexts, evaluating the fluency of languages used only may not validate the quality of the chatbots to be used for cybergrooming prevention programs.
Future Research Directions

1. Conduct deeper data cleaning to find more effective ways to normalize social slangs or online informal languages.

2. Investigate how game theory can optimize the current seq-to-seq model to introduce a perpetrator’s strategic conversations.

3. Develop metrics specific to cybergrooming applications, such as measuring vulnerability or resilience of the languages used in the chatbots.
Acknowledgment & Publications

Acknowledgement

This research is partly supported by Virginia Tech’s ISDA-ISERC (Integrated Security Destination Area-The Integrated Security Education and Research Center) Research Program.

Publications


Any Questions?

Thank You!

Pei Wang
pwang1@vt.edu