SERI: Generative Chatbot Framework for Cybergrooming Prevention

Pei Wang (presenter), Zhen Guo, Lifu Huang, Jin-Hee Cho

Computer Science @ Virginia Tech

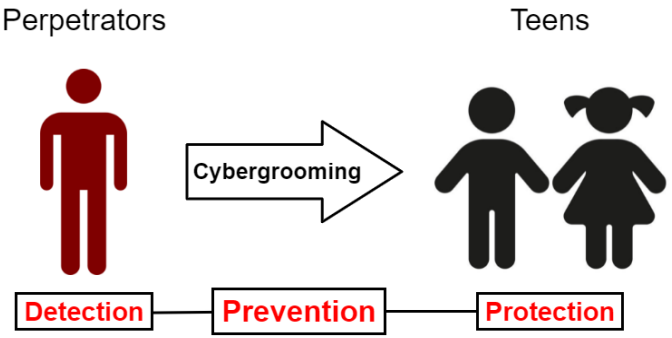
EANCS Workshop, EMNLP 2021, Nov. 11, 2021

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.



Combating 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
- Took a **multi-stage strategies** that the perpetrators take to evolve the relationship with a potential victim.
- Developed a mechanism to **escalate the attack stages** and coordinated the dialogue generation with four subchatbots.

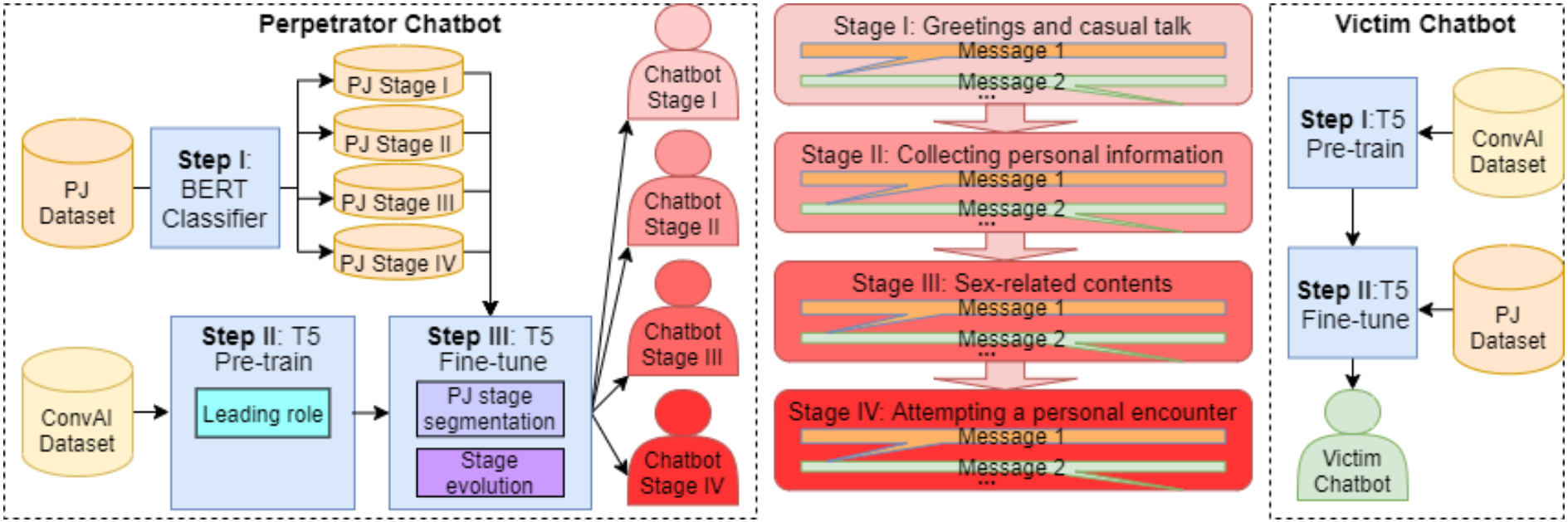
Related Work

- **Detecting 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-based approaches:
  - Negobot (Laorden et al., 2013)
- Pre-training language models:
  - GPT (Radford et al., 2018,2019), BERT (Devlin et al.,2019), and T5 (Raffel et al.,2020)

Limitations:

- Prior effort has focused on the grooming prevention without characterizing the features of victims by cybergrooming.
- No previous research has developed a chatbot to generate conversations between a cybergroomer and a potential victim.

SERI: Architecture Overview



Architecture of the proposed SERI framework

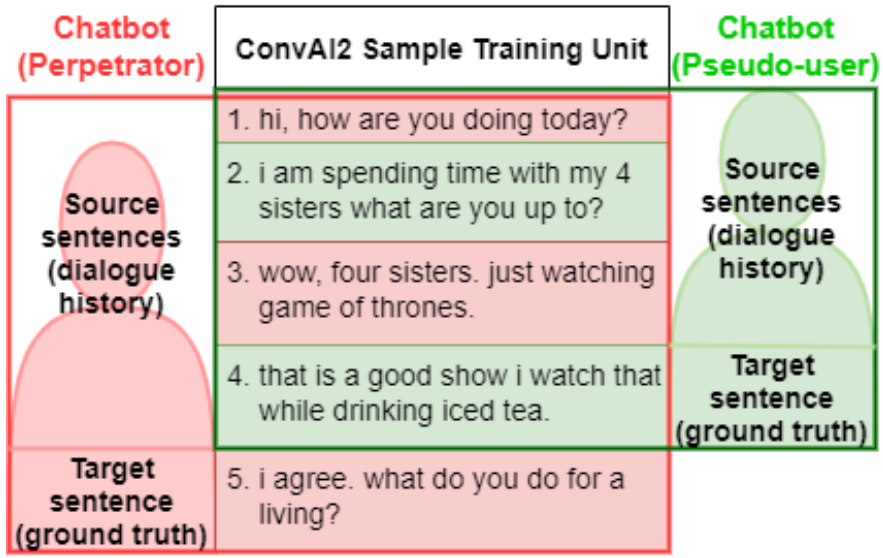
SERI: Cybergrooming Stage Classification

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

- Zambranoet 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 leverage the BERT model to train a stage classifier for the perpetrators.
- We refine the six stages into a new set with four stages.

SERI: Pre-training on the ConvAI2 Dataset

- 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 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 Θ 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.

SERI: Fine-tuning the Chatbots on the PJ Dataset

- **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.

Stages	Label Distribution of Each Block
$\tilde{s}_1$	More than 80% utterances are labeled as $\tilde{s}_1$
$\tilde{s}_2$	More than 60% utterances are labeled as $\tilde{s}_2$
$\tilde{s}_3$	More than 50% utterances are labeled as $\tilde{s}_3$
$\tilde{s}_4$	More than 40% utterances are labeled as $\tilde{s}_4$

Conversation segmentation criteria for the four relationship stages.

SERI: Fine-tuning the Chatbots on the PJ Dataset

- **Chatbot fine-tuning:** We fine-tune the following:
  - The four perpetrator sub-chatbots on the four groups of blocks separately;
  - A victim chatbot based on the victim utterances from the PJ dataset.
- **Output filtering:**
  - Generate 5 candidate messages and select the best one.

	Utterance	Score
Context	hi , how are you doing today ?	
Candidate messages	1. hi, happy your here?, you haven't asked me that yet?	11.14
	2. good, you?	10.03
	3. hi, good you?, you don't need me?	11.14
	4. hi, i'm ok, and you?	**11.17**
	5. ok, how are you?, hello?, hi, hi	10.98

SERI: Stage Evolution of the Perpetrator Subchatbot

- 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 between perpetrator and victim chatbots, the perpetrator will move to stage $\tilde{s}_2$.
- Once the victim detects grooming, he/she will leave the chat conversation immediately and trigger the abortion of conversation.

Experiment Setup

Datasets:

- **The ConvAI2 dataset (Dinan et al., 2019)**
- **The Perverted Justice (PJ) dataset** (PJ Website, 2010)

Metrics:

- **Referenced metrics**: BLEU (Post, 2018), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020)
- **Unreferenced metrics**: MaUde scores (Sinha et al., 2020) and perplexity
- **Human evaluation**: 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, the other generated by the SERI).

Referenced Metrics-based Analysis

Role	BLEU	ROUGE	BERTScore
	Max:100	Max:1	Max:1
Perpetrator	2.9906	0.0970	0.8311
Victim	2.6884	0.1063	0.8274

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 because online chatting languages are often **informal**.
- The BERTScore is relatively high due to:
  - Most of the words are functional and uninformative, which makes it difficult for the BERT to learn meaningful contextual representations;
  - 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

	Perpetrator	Victim
Ground truth dialogues	0.8442	0.8625
Generated dialogues	0.8662	0.8641

MaUde score-based analysis based on PJ evaluation dataset.

- 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

	Perpetrator	Victim
Ground truth dialogues	357.06	477.82
Generated dialogues	139.46	188.97

Perplexity score-based analysis.

- Lower perplexity scores are observed on our generated dialogues.

Human Evaluation Analysis

	Utterance
Context	1: nutting , you miss me 2: ya 3: you better 4: what if i don't ? , lol , jk 5: i'll get you 6: can't get me through the competition duh , i'm not scared of you
Original response	lol, how much you miss me
Generated response	i'm scared of you right now

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).

Conclusions & Future Work

Conclusions:

1. Training the dialogue model with accurate context utterances and target utterance can help distinguish the role of chatbots.
2. Segmenting a training corpus according to each stage can help control the output of a dialogue model.
3. Our human evaluation also shows the promising performance of the SERI by reaching a 37% passing rate for the Turing test.

Future Work Directions:

1. Conduct deeper data cleaning to find more effective ways to normalize social slangs or online informal languages.
2. Investigate how deep reinforcement learning can optimize the current generation model to introduce a perpetrator's strategic conversations.

Any Questions?

Thank You!

Pei Wang
pwang1@vt.edu