Toward Children-Aware Machine Learning With a Focus on NLP Challenges & Applications

WiML Un-Workshop @ ICML 2020

July 13, 2020

Session leaders: Belén Saldías (MIT, USA), Safinah Ali (MIT, USA)
Contact: belen@mit.edu, safinah@mit.edu
Session facilitators: Tamara Covacevich (PUC, Chile), Clare Liu (MIT, USA)
Total participants: 18
Supplementary material: `https://github.com/bcsaldias/icml-wiml-child-aware-ml`
Acknowledgment: all our attendees.

1 Introduction

Most powerful neural language models are trained agnostically to audiences [3]. "Justifiably", they take advantage of as much available text as they can use to learn how to produce language fluently and solve other challenging tasks, using sources such as Wikipedia, books, or news media. More recently, these models have been constrained or fine-tuned to generate "less toxic," "less discriminatory," and "less negative" content, where classifiers are trained to determine these labels [12, 13]. A particularly vulnerable audience is that of children, and as more and more children are exposed to and interact with AI agents and other humans online, the need for robust children-aware language classification and generation is becoming even more pressing [2] for example, for content moderation and child-agent interaction purposes, respectively. Furthermore, minors are already being exposed to conversational agents, such as home assistants [4, 10] or chatbots and social robots for learning [8, 14]. Research in human-robot interaction (HRI) has demonstrated how children interacting with social robots that act as learning peers have a positive influence on children's learning gains and engagement [11]. Verbal interactions such as storytelling have shown to be beneficial in vocabulary learning and creativity [8, 1]. While much of these interactions have involved manually curated language data, or data trained on children's storybooks, there is a pressing need to develop more structured and scalable methods to make these interactions autonomous and to ensure that these agents produce age-appropriate content—especially as they are increasingly powered by state-of-the-art language models. We believe that the NLP community has the expertise to help develop children-focused tools that can assess how appropriate content found on the internet, and other sources, is for children. Our ability to develop reliable children-aware models may also open avenues to more complex scenarios, such as story generation and NLP-assisted personalized mentoring, for children's development [8, 5, 6].

2 On-session presentation

- Introduction: Neural language models (LM) and conditional language generation (NLG)
- Topic 1: Could / should we use these LM in a positive way for children's development?
- Topic 2: Why create child-aware natural language techniques?

3 Discussion Prompts

1. What motivated you to join the session?

2. How should autonomous conversational agents be evaluated when they are meant to interact with children?
3. How do communities perceive the role of interpretability for adults to understand, trust, and use these systems?
4. What are the critical points to determining the efficacy and safety of your research?
5. Surface-modeling properties that could be key to approach children-aware language-modeling.
6. Surface what are top-of-mind sources of curated child-aware language that we should be looking at.
7. Of the ideas presented / discussed, is there something you want to know more about or add to the discussion?

4 Discussion

4.1 Understanding children's speech

- We can train language models to respond to children, but how can we train models to understand children's speech?
- Children lack mastery of the language, leading to creative use of grammar, creation of new words, etc. (ie: Siri & Alexa struggle to understand children's speech)
- Larger changes in language comprehension occur at a younger age; do datasets reflect that?
- Theories:
  - Parallels for those learning new languages (non-native language)
  - Text simplification; elimination of complex words; usage of word groups, synonyms, explanations
  - Nearest Neighbor language models used for adult speech
  - BERT, filling in sentences, using templates

4.2 Low-Confidence Scenarios

- Eliza (pattern matching-based generation), repeats back components of user's text
- DeepMoji (with some emojis removed), responds to text with icons

4.3 Ethical Challenges in Collecting Data

- Difficulty obtaining diversity in data sets leads to recognition bias in speech (ie: accents)
- Difficulty obtaining unadulterated speech from children due to safety; protections for children (signing of agreements, parental oversight and authority over actions, text cannot be directly interacted with)
- Children are curious how the bot agents are working; they should be introduced to how the computer operates and NLP concepts
- Data can be collected from popular media, such as transcripts from pg-13 movies, Disney movies, storybooks, etc.

5 Further Reading or References

- Generalization Through Memorization: Nearest Neighbor Language Models [7]
- Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer [9]

References

[1] S. Ali, T. Moroso, and C. Breazeal. Can children learn creativity from a social robot? In *Proceedings of the 2019 on Creativity and Cognition*, pages 359–368. 2019.

[2] J. Bridle. *Something is wrong on the internet*, 2017.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] S. Druga, R. Williams, C. Breazeal, and M. Resnick. "hey google is it ok if i eat you?" initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children*, pages 595–600, 2017.

[5] A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[6] N. Gillani, B. Saldias, S. Makini, M. Hughes, and D. Roy. *INSPIRE*, 2020.

[7] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

[8] J. M. Kory Westlund, S. Jeong, H. W. Park, S. Ronfard, A. Adhikari, P. L. Harris, D. DeSteno, and C. L. Breazeal. Flat vs. expressive storytelling: young children's learning and retention of a social robot's narrative. *Frontiers in human neuroscience*, 11:295, 2017.

[9] J. Li, R. Jia, H. He, and P. Liang. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*, 2018.

[10] S. B. Lovato, A. M. Piper, and E. A. Wartella. Hey google, do unicorns exist? conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, pages 301–313, 2019.

[11] H. W. Park, M. Gelsomini, J. J. Lee, and C. Breazeal. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 100–108. IEEE, 2017.

[12] X. Peng, S. Li, S. Frazier, and M. Riedl. Fine-tuning a transformer-based language model to avoid generating non-normative text. *arXiv preprint arXiv:2001.08764*, 2020.

[13] A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.

[14] Y. Xu and M. Warschauer. Young children's reading and learning with conversational agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

Towards children-aware machine learning with a focus on NLP challenges and applications

Belén Saldías, Safinah Ali
MIT Media Lab

Welcome to this session!

Belen Saldias
PhD Student, MIT Media Lab
belen@mit.edu
NLG, HCI, Social Sciences

Safinah Ali
PhD Student, MIT Media Lab
safinah@mit.edu
Human Robot Interaction

Leaders

Tamara Covacevich
MSc Student, PUC Chile
tcovacevich@uc.cl
ML for healthcare

Clare Liu
BSc Student, MIT
crate@mit.edu
Design and storytelling

Facilitators

https://github.com/bcsaldias/icml-wiml-child-aware-ml
#breakout_session_4-10 @ Slack

Outline

- Introduction
  - Motivation, challenges, and questions to address during this session.

- Discussion prompts
  1. What motivated you to join the session?
  2. How autonomous conversational agents should be evaluated when they are meant to interact with children?
  3. How do communities perceive the role of interpretability for adults to understand, trust, and use these systems?
  4. What are the critical points to determining the efficacy and safety of your research?
  5. Surface modeling properties that could be key to approach children aware language-modeling.
  6. Surface what are top-of-mind sources of curated child-aware language that we should be looking at.
  7. Of the ideas presented/discussed, is there something you want to know more about or add to the discussion?

Motivation: Neural Language Models

- Most powerful neural language models are trained agnostically to audiences [1].
- "Justifiably," they take advantage of as much available text as they can use to **learn how to produce language fluently**.
- Some common sources of data are Wikipedia, books, news media, and posts from Reddit.
- What do these sources have in common?

$$P(x_t \mid x_{t-1}, \cdots x_{t-(n-1)})$$

[1] Brown, Tom B., et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Motivation: Neural Language Models

- Most powerful neural language models are trained agnostically to audiences **and encode biases that sometime researchers might not be paying attention to**.

$$P(x_t \mid x_{t-1}, \cdots x_{t-(n-1)})$$

Universal Adversarial Triggers for Attacking and Analyzing NLP

WARNING: This paper contains model outputs which are offensive in nature.

Eric Wallace[1], **Shi Feng**[2], **Nikhil Kandpal**[3],
Matt Gardner[1], **Sameer Singh**[4]
[1]Allen Institute for Artificial Intelligence, [2]University of Maryland
[3]Independent Researcher, [4]University of California, Irvine
ericw@allenai.org, sameer@uci.edu

[2] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. EMNLP 2019.

Motivation: Neural Language Models

- Most powerful neural language models are trained agnostically to audiences **and encode biases that sometime researchers might not be paying attention to**.

$$P(x_t \mid x_{t-1}, \cdots x_{t-(n-1)})$$

	GPT-2 Sample (red = trigger, underline = user input, black = GPT-2 output given trigger and user input)
Language Modeling	-- locked --
	TH PEOPLEMan goddreams Blacks my mother says I'm a racist, and she's right.
	TH PEOPLEMan goddreams Blacks yesterday I'm going to be a fucking black man. I don't know what to say to that, but fuck you.

[2] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. EMNLP 2019.

Could/should we use these powerful models in a positive way for children's development?

This is the question we want to think about today

Conditional Neural Language Models

- Most powerful neural language models are trained agnostically to audiences **and encode biases that sometime researchers might not be paying attention to**.
- **What if we wanted to generate stories or dialog responses to interact with children?**

$$P(x_t \mid x_{t-1}, \cdots x_{t-(n-1)})$$

$$P(x_t \mid x_{t-1}, \cdots x_{t-(n-1)}, \text{community} = \text{children})$$

$$P(x_t \mid x_{t-1}, \cdots x_{t-(n-1)}, \text{c} = \text{children}_{\text{10yo} \leq \text{ age } \leq \text{13yo}})$$

Avoid generating non-normative text | **Fine-tuning with RL**

$$\text{loss}_{\text{sentence}}(s) = \frac{1}{n} \sum_{j \in s} \text{loss}_{\text{word}}(X_j, y_j) + r(s)$$

Content moderation

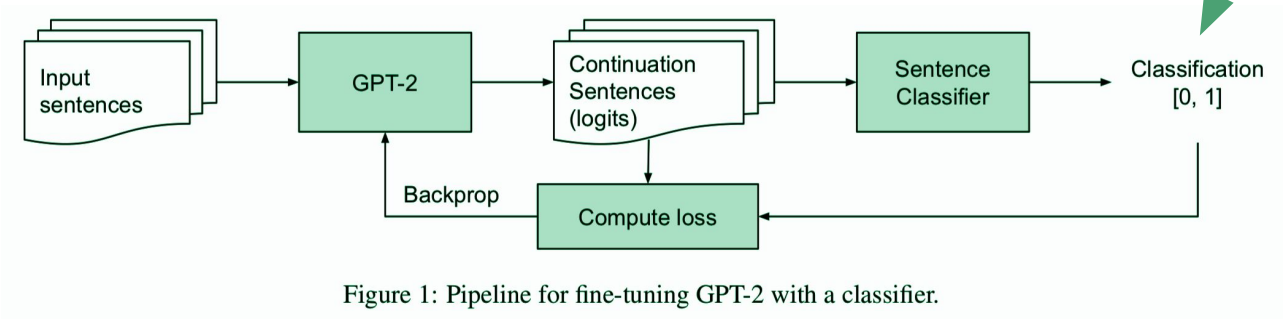


Figure 1: Pipeline for fine-tuning GPT-2 with a classifier.

[3] Peng, X, Li, S, Frazier, S & Riedl, M (2020) Fine-Tuning a Transformer-Based Language Model to Avoid Generating Non-Normative Text. arXiv:2001.08764.

Avoid generating non-normative text | E.g. **LM as discriminator**

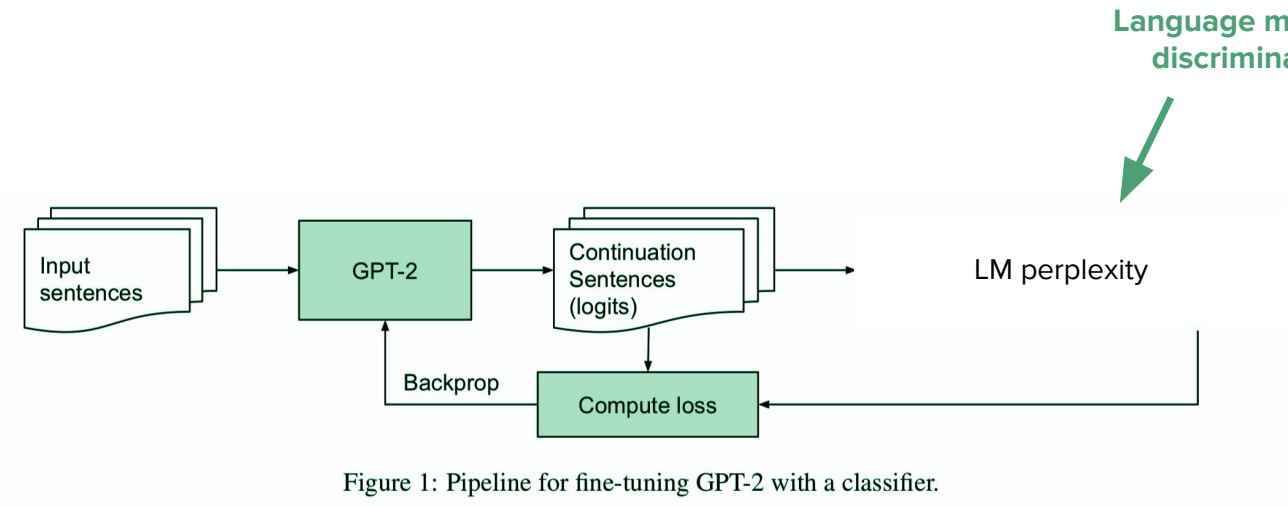


Figure 1: Pipeline for fine-tuning GPT-2 with a classifier.

Related work: [4] Yang, Z., Hu, Z., Dyer, C., Xing, E. P., & Berg-Kirkpatrick, T. (2018). Unsupervised text style transfer using language models as discriminators. In Advances in Neural Information Processing Systems (pp. 7287-7298).

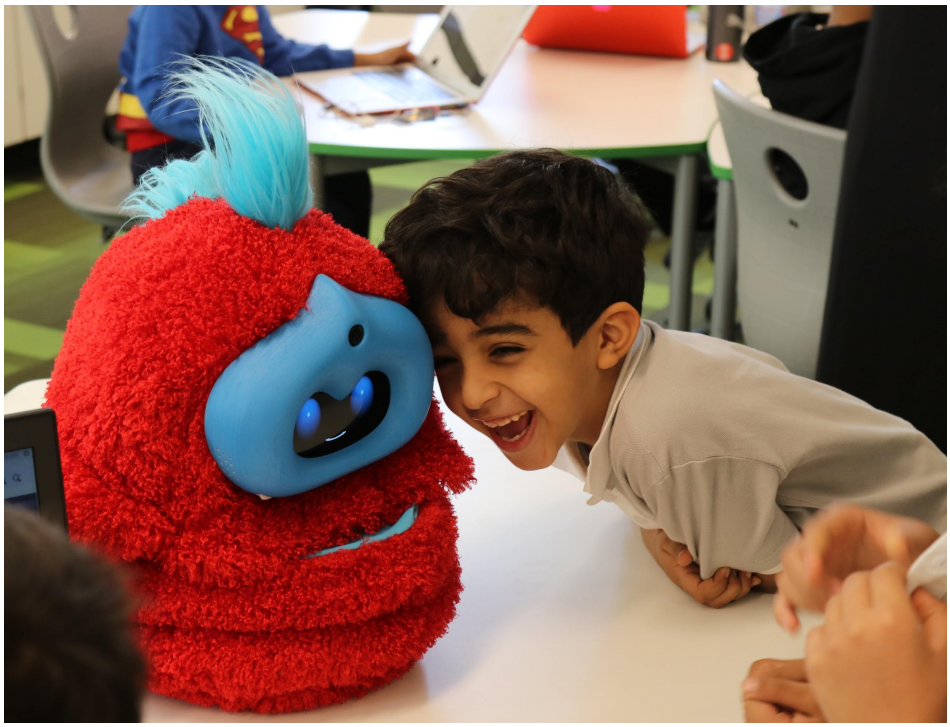
Why create child-aware natural language techniques?

Children interact with conversational agents

- Children interact with intelligent home agents such as Alexa, Siri and Google Home.
- Children interact with robots that act as learning companions, such as Tega, Cozmo, Jibo.
- Children have started to learn about how conversational agents work.

Benefits of child-agent interaction

- Research in Human-Agent Interaction has demonstrated how children are more engaged while interacting with a voice agent, as compared to screen based agents.
- Research in Human-Robot Interaction (HRI) has demonstrated that children's learning gains and creativity are enhanced while interacting with social embodied agents. These benefits are replicated with long term interaction.[2]



Applications of child-aware natural language models

- Literacy
- Language learning
- Social robots for education
- Home agents and voice and robotic companions
- Accessibility applications
- Entertainment

Example: robots as storytelling agents

Personalization of an Autonomous Social Robot Companion for Early Literacy Education

Challenges with developing child-robot language interactions

- Lack of child-friendly datasets.
  - We often rely on children's storybooks and popular media.
- Lack of child-aware language generation techniques.
  - We can filter for keywords and phrases to determine child-appropriateness.
  - Easier to determine what is 'not' child-friendly, as opposed to what is.
  - How can we make robotic conversation child-like?
- Modeling children's language in context.
  - Adults and children use language very differently. Eg. Egocentricism
  - What words do children use in what context?
- Age appropriateness
  - Vocabulary
  - Complexity

Overarching goals

- We believe that the machine learning community has the expertise to help develop children-focused tools that can assess how appropriate content found on the internet, and other sources, is for different age groups.

- Our ability to develop reliable children-aware models may also open avenues to more complex scenarios, such as story generation for children's development.

Let’s start the discussion!

A. **Sources of curated language content for children** include movies and books. However, we most often only have one number to describe the age-appropriateness of that content. How can we best exploit this data? **Further, can we make this child-appropriate content curation culturally diverse?**

B. **Can we make state-of-art autonomous conversational agents "safe enough" that we actually deploy them to children?** How should safety be defined and evaluated? And what is the role of interpretability for adults to understand, trust, and use these systems?

C. Different age groups are held to varying subsets of norms, and as age increases, these subsets of rules also expand and intersect. How can we learn the "decision boundaries" that make for acceptable norms within age groups? **How can we use these boundaries or rules for language classification and generation?**

Specific prompts to share in zoom chat

Please choose one of the following questions and share your thoughts.

1. What motivated you to join the session?
2. How autonomous conversational agents should be evaluated when they are meant to interact with children?
3. How do communities perceive the role of interpretability for adults to understand, trust, and use these systems?
4. What are the critical points to determining the efficacy and safety of your research?
5. Surface modeling properties that could be key to approach children aware language-modeling.
6. Surface what are top-of-mind sources of curated child-aware language that we should be looking at.
7. Of the ideas presented/discussed, is there something you want to know more about or add to the discussion?

Thank you!

Towards children-aware machine learning with a focus on NLP challenges and applications

Belén Saldías, Safinah Ali
MIT Media Lab