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

Belen Saldias and Safinah Ali
MIT Media Lab
{belen, safinah}@mit.edu

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 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 [2, 3]. 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 [4]: for example, for content moderation and child-agent interaction purposes, respectively. Furthermore, minors are already being exposed to conversational agents, such as home assistants [5, 6] or chatbots and social robots for learning [7, 8]. 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 [9]. Verbal interactions such as storytelling have shown to be beneficial in vocabulary learning and creativity [7, 10]. 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 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 [7, 11, 12].

Some of the challenges that we would like to discuss concerning what children-aware NLP entails are:

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? [2, 3]

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? [13]

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? [14, 15]

Format: We will spend the first 10 minutes introducing the topic, and end with a slide of questions to discuss. We will then ask attendees to introduce themselves, pick a question to give an opinion on, share how what they do aligns with this area, and promote discussion from there. We will wrap up by inviting people to explore next steps.

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., ... & Agarwal, S. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.
2. 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*.
3. Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. *In Proceedings of the Fifth International Workshop on NLP for Social Media*.
4. Something is wrong on the internet (2017). Retrieved 8 June 2020 from *medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2*
5. Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017, June). "Hey Google is it OK if I eat you?" Initial Explorations in Child-Agent Interaction. *Conference on Interaction Design and Children*.
6. Lovato, S. B., Piper, A. M., & Wartella, E. A. (2019, June). Hey Google, Do Unicorns Exist? Conversational Agents as a Path to Answers to Children's Questions. *International Conference on Interaction Design and Children*.
7. Kory Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., ... & Breazeal, C. L. (2017). Flat vs. expressive storytelling: young children's learning and retention of a social robot's narrative. Frontiers in human neuroscience, 11, 295.
8. Xu, Y., & Warschauer, M. (2019, May). Young children's reading and learning with conversational agents. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-8).
9. H. W. Park, M. Gelsomini, J. J. Lee and C. Breazeal, "Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling *," 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, Vienna, 2017, pp. 100-108.
10. Ali, S., Moroso, T., & Breazeal, C. (2019). Can Children Learn Creativity from a Social Robot?. In Proceedings of the 2019 on Creativity and Cognition.
11. Cremin, T., Flewitt, R., Mardell, B, & Swann, J (2016) Storytelling in early childhood: Enriching language, literacy and classroom culture. Taylor & Francis
12. Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. arXiv preprint arXiv:1805.04833.
13. Jacovi, A., & Goldberg, Y. (2020). Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?. arXiv:2004.03685.
14. Awasthi, A., Ghosh, S., Goyal, R., & Sarawagi, S. (2020) Learning from Rules Generalizing Labeled Exemplars. ICLR.
15. Dathathri, S., Madotto, A, Lan, J., ... & Liu, R. (2019) Plug and play language models: a simple approach to controlled text generation. *arXiv:1912.02164*.