Robust metrics for the evaluation of fairness and safety in NLG

Johnny Tian-Zheng Wei
Natural Language Processing

Motivation. As language technology matures, we can expect natural language generation (NLG) systems to be more pervasive in our daily life. Companies routinely make headlines on successfully applying NLG to settings ranging from news personalization, customer service, and social media marketing. Startups such as Replika even specialize in developing chatbot technology to alleviate lonliness [1]. The rise in NLG applications is in part due to recent developments of end-to-end, data-driven neural generation models. These are expressive models that can learn to generate arbitrary sequences of words, and can be learned largely from data, significantly reducing engineering effort. As an example, it is now now easier than ever to prototype data-driven, state-of-the-art dialog models using publicly available conversational datasets and dialog-specific pretrained models trained on massive amounts of web conversations [2].

Incorporating neural generation techniques that leverage large amounts of data into NLG empirically gives us better performance. However, we must recognize that our language found in the training data takes on social meaning. For instance, stereotypes, reflected in language, serve to spread beliefs about social groups to reinforce hierarchies, leading to social inequality. In extreme cases, language incites behavior such as self-harm, hate, or violence. It is known that **NLG systems can replicate, or even amplify the toxicity found in their training data [3].** As a commercial example, Tay [4], which Microsoft released to Twitter in 2016, tweeted racist and sexually-charged messages learned from its training data. For startups such as Replika, an insensitive utterance to a vulnerable individual could have severe consequences. ***Robust* identification of unfair and unsafe language is necessary in deployment and crucial to mitigation in training time** to ensure our systems do not perpetuate societal inequalities or harm at-risk individuals.

The methods to determine fairness and safety in NLG is converging on "output, classify, and count" - where 1) NLG systems are prompted for a set of outputs, then 2) an automatic classifier labels each output e.g. whether the output contains a microaggression, and finally 3) the number of positive labels are reported [5]. For language generation, the use of an *automatic* classifier is necessary, as the amount of equivalent human annotation implicated may be impractical in development cycles. In the current paradigm, the introduction of an imperfect classifier is problematic, as the errors of the classifier should be taken into account for scientifically valid inference. This proposal seeks to address two critical robustness issues with current fairness and safety metrics in NLG deployment, and propose a mitigation technique based on robust metrics in training:

(1) Bridging the gap between metrics and quantification. Currently metrics research originating from the NLG communities does not engage with quantification research, which has its roots in statistics and social science. Fundamentally, both fields count with classifiers. These fields are complementary - quantification provides the statistical tools to adjust the observed counts based on classifier errors, and metrics provide the underlying models that can be used to detect the often complex unfair or unsafe phenomenon. There are two directions we can take to connect the two fields: 1) calibrating metric output probabilities with techniques such as Platt scaling so they may interplay with existing quantification tools [6], or 2) adopt structured losses derived from group-level constraints in training metrics, with the goal to equalize false positive and false negative rates so metric errors may cancel out when used for counting [7].

(2) Training metrics robust to confounding features. For the underlying classifier used in quantification, an assumption on its generalization breaks down when there are confounding features

(even when adjusting for error rates as in **(1)** the classifier's error rates may be unstable in the presence of confounders). There is precedent that a hate-speech classifier's false positive rate increases in the presence of African-American dialectical features [8]. If we were to use a hate-speech classifier confounded by topics, we will not be able to accurately conclude which topics dialog systems are more predisposed to generate hate-speech. To combat confounders, I will apply adversarial training for metrics to maximize accuracy while minimizing the accuracy of an adversary predicting the presence of confounders with the same features. Previous work shows that adversarial techniques can reduce the false-positive rates while minimally affecting accuracy [8].

(3) Mitigating fairness and safety risks with robust metrics. With minimum risk training [MRT;9], it is possible to optimize a neural NLG system with respect to metrics such as the ones we proposed in **(1)** and **(2)** to mitigate risks. Naturally, we may ask whether mitigation of risks in NLG systems with MRT is possible with our robust metrics. However, straightforward application is known not to work, as NLG systems may learn degenerate outputs to optimize the metric, or are unable to learn effectively from metric signal [10]. This final proposal seeks to address these open-ended engineering challenges by 1) identifying and integrating the smoothness constraints for stable MRT training, and 2) reduce the possibilities of NLG systems to exploit the metric in training by applying techniques for defenses against adversarial attacks to the metric.

Resources. Our research will first build on prior work on offensive speech in dialog, however the techniques proposed are generalizable to any definition of fairness and safety in NLG.

Intellectual merit. The current "output, classify, and count" paradigm of fairness and safety in NLG introduces an imperfect classifier which raises validity concerns. This proposal seeks to address these critical concerns, and uses the resulting robust metrics for mitigating risks in NLG systems during training. These concerns also hold in general NLG evaluation, and so the results from this proposals will also directly benefit NLG evaluation. Furthermore, the robust classification issues discussed here extend beyond fairness and safety metrics, and are isomorphic to classification bias issues in general. Finally, optimizing NLG systems with learned metrics is also a major research direction within NLG.

Broader impact. Our society is consuming more and more generated text. Pessimistically, NLG systems have the potential to harm vulnerable members of our society by learning from the language humans developed to harm each other. In addition, representational harms or the use of microaggressions can contribute to societal inequality at large, which is also far from desirable. By improving the robustness of fairness and safety metrics through the issues proposed here, we actively neutralize language that reinforces societal inequality, and will be better equipped to deploy fairer and safer NLG systems to broad audiences, for the benefit of all.

---

[1] https://replika.ai/ [2] Zhang et al., "DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation" in *ACL (demo) 2020*. [3] Gehman et al., "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models" in *EMNLP Findings 2020*. [4] https://en.wikipedia.org/wiki/Tay_(bot) [5] Liu et al., "Does Gender Matter? Towards Fairness in Dialogue Systems" in *COLING 2020*. [6] Card et al., "The importance of calibration for estimating proportions from annotations" in *NAACL 2018*. [7] Zhao et al., "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints" in *EMNLP 2017*. [8] Xia et al., "Demoting Racial Bias in Hate Speech Detection" in the *SocialNLP Workshop @ ACL 2020*. [9] Shen et al., "Minimum Risk Training for Neural Machine Translation" in *ACL 2016*. [10] Weiting et al., "Beyond BLEU: Training Neural Machine Translation with Semantic Similarity" *ACL 2019*.