MESURER LES BIAIS STÉRÉOTYPÉS DANS LES MODÈLES DE LANGUE AUTO-REGRESSIFS

EVALUATING STEREOTYPED BIASES IN AUTO-REGRESSIVE LANGUAGE MODELS

Etablissement **Université de Lorraine**

École doctorale **IAEM - INFORMATIQUE - AUTOMATIQUE - ELECTRONIQUE - ELECTROTECHNIQUE - MATHEMATIQUES**

Spécialité **Informatique**

Unité de recherche **LORIA - Laboratoire Lorrain de Recherche en Informatique et ses Applications**

Encadrement de la thèse **Karën FORT (detailResp.pl?resp=94970)**

Financement *Employeur* **Université de Lorraine**

Début de la thèse le **1 octobre 2023**

Date limite de candidature (à 23h59) **15 mai 2023**

Mots clés - Keywords

biais, éthique

biases, ethics

Description de la problématique de recherche - Project description

L'objectif de la recherche doctorale est de fournir une compréhension fine des biais encodés dans les modèles de langage auto-régressifs. Plus précisément, le ou la doctorant·e produira des ressources et des outils pour l'évaluation extrinsèque des stéréotypes et mènera une évaluation complète des modèles de langue qui englobe une dimension éthique ainsi que des mesures de performance.

Une première étape du travail consistera à construire un état de l'art solide sur l'évaluation des biais stéréotypés. Cela devrait inclure toutes les méthodes extrinsèques, y compris l'ingénierie de prompt, ainsi que les mesures existantes.

En parallèle, le ou la doctorant·e déterminera si des jeux de données précédemment créés, tels que CrowS-Pairs (Nangia2020) et ses adaptations dans d'autres langages comme le French CrowS-Pairs (Neveol2022) peuvent être réutilisés dans le cadre de modèles de langue auto-régressifs et proposer des métriques adaptées.

Une autre dimension que nous voulons couvrir dans ce travail est de vérifier la cohérence des résultats obtenus sur les applications de pré-apprentissage des modèles (par exemple avec CrowS-Pairs) et certaines applications plus en aval. Les candidats potentiels pourraient être des applications de TAL médical, telles que l'extraction d'indicateurs épidémiologiques à partir de récits cliniques.

L'ensemble des ressources et outils produits seront mis à disposition de l'ensemble de la communauté sur un dépôt en ligne librement accessible (Inria GitLab).

The objective of the doctoral research is to provide a fine-grained understanding of biases encoded in auto-regressive language models. Specifically, the PhD candidate will produce resources and tools for the extrinsic evaluation of stereotyped biases and conduct a comprehensive evaluation of language models that encompasses an ethical dimension as well as performance metrics.

A first step of the work will consist in building a solid state-of-the-art about stereotyped biases evaluation. This should include all extrinsic methods, including prompt engineering, as well as the existing metrics.

In parallel, the PhD candidate will determine if previously created datasets, such as CrowS-Pairs (Nangia2020) and its adaptations in other languages like French CrowS-Pairs (Neveol2022) can be re-used in the context of auto-regressive language models and propose appropriate metrics.

Another dimension that we want to cover in the work is to check the consistency of the results obtained on the models' pre-training applications (eg with CrowS-Pairs) and some more downstream applications. Potential candidates could be NLP applications supporting public health, such as the extraction of epidemiological indicators from clinical narratives, as we have experience on these.

All the produced resources and tools will be made available to the entire community on a freely accessible online repository (Inria GitLab).

Thématique / Contexte

Large language models have been at the heart of the majority of Natural Language Processing (NLP) tools and applications for the past 4 years now. After the success of the masked language models (eg BERT), auto-regressive language models (eg GPT) are now the most widely used. However, regardless of the architecture and the languages they cover, these models reproduce and amplify the stereotypes which are present in the datasets used to train them (Zhao2017, Jia2020). These stereotyped biases have a strong negative impact on society, especially on the historically most disadvantaged groups (Hovy2016, Bender2021).

Many research efforts have focused on mitigating these negative effects, either by improving the documentation of the corpora used for training (Couillault2014, Bender2018, Gebru2021) or by debiaising the models (Meade2022). Other efforts aim at producing specific data allowing to measure the degree of stereotype of the productions (Nangia2020, Neveol2022).
However, these experiments mainly addressed the masked language models (like BERT (Devlin2019), not the auto-regressive ones like GPT3 (Brown2020) or Bloom (Scao2022). With the advent of chatGPT, a variant of auto-regressive model using Reinforcement Learning from Human Feedback (RLHF), and the numerous issues uncovered by the users1, the urge for a scientifically sound methodology of evaluation has become obvious.
Finally, most research work in bias and fairness in NLP is focused on gender bias in American English (Talat et al. 2022).

Références bibliographiques

[Bender and Friedman, 2018] Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing : Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6 :587–604.
[Bender et al., 2021] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots : Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.
[Blodgett et al., 2020] Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of 'bias' in nlp. In ACL.
[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.
[Couillault et al., 2014] Couillault, A., Fort, K., Adda, G., and De Mazancourt, H. (2014). Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter. In International Conference on Language Resources and Evaluation (LREC), Reykjavik, Islande.
[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
[Gebru et al., 2021] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. (2021). Datasheets for datasets. Commun. ACM, 64(12) :86–92.
[Hovy and Spruit, 2016] Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers), pages 591–598, Berlin, Germany. Association for Computational Linguistics.
[Jia et al., 2020] Jia, S., Meng, T., Zhao, J., and Chang, K.-W. (2020). Mitigating gender bias amplification in distribution by posterior regularization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2936–2942, Online. Association for Computational Linguistics.
[Meade et al., 2022] Meade, N., Poole-Dayan, E., and Reddy, S. (2022). An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.
[Mitchell et al., 2019] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery. 2
[Nangia et al., 2020] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

[Névéol et al., 2022] Névéol, A., Dupont, Y., Bezançon, J., and Fort, K. (2022). French crows-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Irlande.

[Ouyang et al., 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, Advances in Neural Information Processing Systems.

[Scao et al., 2022] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., et al. (2022). BLOOM : A 176B-Parameter Open-Access Multilingual Language Model. working paper or preprint.

[Talat et al., 2022] Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

[Zhao et al., 2017] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping : Reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics

Précisions sur l'encadrement - Details on the thesis supervision

Aurélie Névéol (co-directrice, DR CNRS, LISN)

Conditions scientifiques matérielles et financières du projet de recherche

La recherche aura lieu au LORIA, à Nancy, au sein de l'équipe Sémagramme.

Objectifs de valorisation des travaux de recherche du doctorant : diffusion, publication et confidentialité, droit à la propriété intellectuelle,...

Les travaux de recherche feront l'objet de publications régulières en accès libre. Toutes les données produites seront mises à la disposition de la communauté sous licence CC sur une plateforme en ligne. Le code produit sera également mis à disposition sous licence libre sur une plateforme.

Profil et compétences recherchées - Profile and skills required

Master en Traitement automatique des langues

Intérêt envers les sujets de l'éthique et de la création de données

Excellent niveau de français, très bon niveau d'anglais.
Required qualifications:

MSc in Natural Language Processing.
Interest in ethics for NLP and datasets building.

Languages:
Fluent French and very good English.

Dernière mise à jour le 8 avril 2023