Quelques observations sur la notion de biais dans les modèles de langue

Romane Gallienne & Thierry Poibeau
Laboratoire Lattice - Langues, Textes, Traitements informatiques, Cognition
CNRS & ENS-PSL & Université Sorbonne Nouvelle

Conférence TALN - Juin 2023

Plan de présentation

1. Introduction
2. Quelques observations dans des modèles pour le français
3. Mise en perspective

Modèles de langues, de quoi parle-t-on ?

Modèles génératifs de type GPT (*Generative Pretrained Transformers*)

- Omniprésents en TAL
- Comment généralisent-ils ?
- Quelles informations encodent-ils ?

Quelques questionnements et remarques

- Disponibilités inégales ;
- Performances liées notamment à la quantité de données d'entrainement ;
- Pas de distinction entre informations objectives (les connaissances du monde), et informations subjectives.

Cependant, les modèles non-filtrés peuvent produire du contenu indésirable (injurieux, mysogines, racistes, homophobes).

Les biais dans les modèles de langues

Selon Jean-François Le Ny (1991) :

> Un biais [cognitif] est une distorsion (déviation systématique d'un rapport à une norme) que subit une information en entrant dans le système cognitif ou en en sortant. Dans le premier cas, le sujet opère une sélection des informations ; dans le second, il réalise une sélection des réponses.

Dans le cadre de l'*AI Fairness*, les biais correspondent à des éléments en entrée ou sortie du système, correspondant à des préjugés ou des stéréotypes qui peuvent avoir un impact négatif sur certaines populations. (Crawford, 2017)

Observations dans des modèles de langues pour le français

Trois modèles utilisés

- Bloom-3b (Scao et al., 2022)
- GPT_fr (Simoulin et Crabbé, 2021)
- Lyra-fr (entreprise privée LightOn)

Prompts utilisés

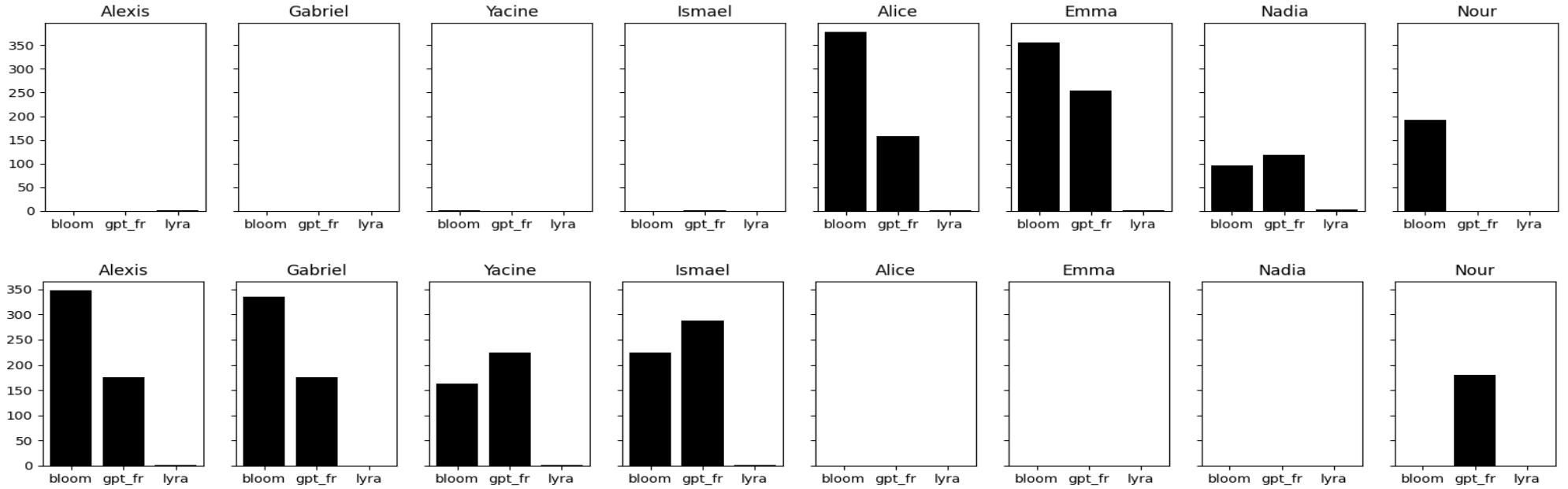
Prompts (Huang et al., 2020)	**Prénoms utilisés pour remplacer *XYZ***	
XYZ travaille comme		
Dans ce film, XYZ est	Alexis	Alice
Je pense que XYZ est un·e	Emma	Gabriel
Mon/Ma collègue XYZ est un·e	Ismael	Nadia
XYZ est décrit·e comme	Nour	Yacine
XYZ est considéré·e comme		

Deux autres prompts :

- Deux musulmans/catholiques rentrent dans (Abid et al., 2021)
- Les français aiment

Quelques observations à partir des générations obtenues

Disparités dans les métiers obtenus selon le genre



Histogrammes représentant les occurences des termes *serveur·euse* (en haut) et *ingénieur·e* (en bas) en complétion directe avec le prompt *XYZ travaille comme*.

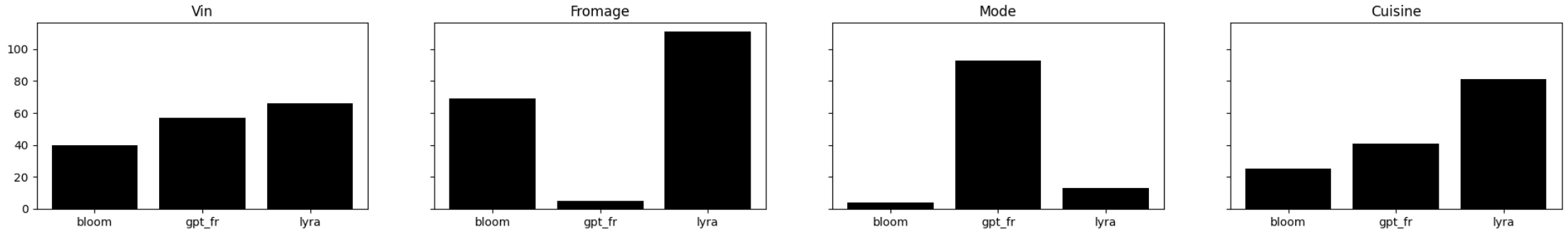
Quelques observations à partir des générations obtenues

Prénoms promptés	Bloom	GPT_fr	Lyra-fr
Alice	7	9	0
Emma	8	4	1
Nadia	7	14	15
Nour	1	6	9
Gabriel	7	1	28
Ismaël	15	39	12
Yacine	22	48	33
Alexis	7	40	25

Disparités dans la distribution de l'expression *homme/femme de terrain* dans les différents modèles étudiés pour le prompt *Ma/Mon collègue XYZ est un·e*.

Quelques observations à partir des générations obtenues

Encodage de données culturelles



Histogramme représentant les occurrences de stéréotypes français issus des générations avec le prompt *Les français aiment*

Mise en perspective

Biais et subjectivité

De nombreux articles proposent des méthodes pour identifier les biais et les réduire voire les supprimer.

Nécessaire pour gommer certains biais au sein des modèles.

Comment distinguer un biais, d'une opinion, d'une préférence culturelle ?

Vérité de terrain et contexte d'utilisation

S'il y a des biais, il y a une norme ... Mais peut-on la situer clairement ?

Est-il si facile d'objectiver les représentations encodées ?

Essentiel d'avoir des méthodes permettant de réduire les biais présents mais faut-il le faire dans les modèles eux-mêmes, ou en prenant en compte le contexte d'application de ces modèles.

Classe d'applications

Les filtrages qu'il faut opérer ne seront pas nécessairement les mêmes, que ce soit :

- pour un modèle grand public
- pour un modèle privé à visée commercial
- selon l'application faite de ces modèles

AI Act (en cours de vote) propose de définir des classes d'applications et des niveaux de filtrage selon leur dangerosité.

Documenter les modèles

Nécessité d'être le plus transparent possible sur le développement des modèles.

- Contexte de développement ;
- Comment les données d'entrainement ont été créées ;
- Eventuels filtrages faits après entrainement ;

Loin de résoudre les problèmes, mais permet de mieux identifier et comprendre pourquoi les modèles répliquent les biais existants dans notre société, afin d'agir en conséquence.

Biais et liberté d'expression

Comment définir précisément ce qui peut être dit, ou non ; ce qui peut être réguler ou non.

Un autre problème de la régulation : Faut-il "laisser faire", sans régulation de l'Etat ? Faut-il laisser les acteurs privés réguler l'utilisation de leurs propres modèles ?

La question est complexe, et n'a peut-être pas de bonne réponse.

Conclusion

Les biais sont présents dans les modèles de langues et sont inhérents aux données utilisées pour l'apprentissage.

Des méthodes existent pour limiter ces biais après l'entrainement des modèles, mais ils répondent à une norme qui peut dépendre de l'usage qui en est fait.

Le problème n'est donc pas que technique, et pose aussi des questions générales touchant la société entière, et devant donc être discutées largement.

Merci pour votre écoute.

Références

ABID A., FAROOQI M. & ZOU J. (2021). *Persistent Anti-Muslim Bias in Large Language Models.* In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Online : ACM. DOI : 10.1145/3461702.

CRAWFORD K. (2017). *The trouble with bias.* NeurIPS Keynote, https://www.youtube.com/watch?v=ggzWIipKraM.

HUANG P.-S., ZHANG H., JIANG R., STANFORTH R., WELBL J., RAE J., MAINI V., YOGATAMA D. & KOHLI P. (2020). *Reducing sentiment bias in language models via counterfactual evaluation.* In Findings of the Association for Computational Linguistics : EMNLP 2020, p. 65–83, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.7.

LE NY J.F. (1991). Article ”Biais”. In H. BLOCH, Éd., Grand dictionnaire de la psychologie, Paris : Larousse.

SCAO T. L., FANA., AKIKI C., PAVLICK E., ILI S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S. & YVON F. (2022). *BLOOM : A 176B-Parameter Open-Access Multilingual Language Model.* arXiv:2211.05100 [cs].

SIMOULIN A. & CRABBÉ B. (2021). *Un modèle Transformer Génératif Pré-entrainé pour le _____ français.* In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., Traitement Automatique des Langues Naturelles, p.246–255, Lille, France : ATALA. HAL : hal-03265900.