La linguistique appliquée pour une IA plus éthique

Fanny Ducel*[1], Karën Fort[2], and Aurélie Névéol[3]

[1]Laboratoire Interdisciplinaire des Sciences du Numérique – Université Paris-Saclay, Centre National de la Recherche Scientifique – France
[2]SEMAGRAMME (INRIA Nancy - Grand Est / LORIA) – Université de Lorraine, INRIA – France
[3]Laboratoire Interdisciplinaire des Sciences du Numérique – Université Paris-Saclay, Centre National de la Recherche Scientifique – France

Résumé

Le Traitement Automatique des Langues et l'IA connaissent un engouement scientifique et médiatique énorme, notamment avec l'avènement des modèles de langues (LM). De nouveaux LM sont mis en ligne quotidiennement et sont supposément de plus en plus performants. Néanmoins, la linguistique et l'éthique sont laissées pour compte. Nous avançons que ces deux disciplines ont un rôle crucial à jouer pour les LM, et encourageons les linguistes à s'emparer des objets d'études que sont les modèles de langues et leurs productions.
Tout d'abord, la linguistique appliquée pourrait permettre d'identifier les limites des LM. En effet, on sait que les LM sont entraînés sur des masses de données trouvées sur le Web (contenu issus de réseaux sociaux, articles de presse, œuvres littéraires, ...). Au-delà du problème de droits d'auteurs bafoués, cet entraînement implique le gel d'une langue et la diminution de la variation : c'est une version d'une langue précise, à un instant donné dans le temps, telle qu'utilisée par un certain groupe de locuteur·ices (par exemple, la majorité des personnes qui ont écrivent sur Wikipedia sont des hommes occidentaux qui ont la vingtaine ou sont retraités 1) qui sera apprise, reproduite et amplifiée par les LM. Les productions des LM sont alors des objets d'étude linguistique pertinents, et leur analyse permettrait notamment de documenter ce gel (Hovy et al. (2020)).

La linguistique appliquée pourrait également servir à détecter les biais stéréotypés qui sont reproduits et amplifiés par l'IA (Jia et al. (2020)). Beaucoup d'études visent à identifier et diminuer ces biais, mais la plupart s'appuient elles-mêmes sur de l'IA ou sur des listes de mots établies manuellement. (Ducel et al. (2024)) optent pour une approche basée sur des règles et ressources linguistiques afin de développer un système de détection du genre en français et en italien, et pouvoir ainsi évaluer automatiquement les biais de genre dans des générations. Ce type de systèmes est plus transparent que l'IA (effet " boîte noire ") et bien moins coûteux du point de vue environnemental. Ils nécessitent moins de matériel et de temps de calcul, ce qui les rend également plus accessibles à des communautés scientifiques et linguistiques qui disposent de peu de moyens et sont victimes des biais des LMs.

Par ailleurs, la linguistique appliquée pourrait aider à déceler des biais plus implicites, à l'aide d'analyses lexicales, syntaxiques ou sémantiques plus détaillées, mais aussi à tendre vers une plus grande inclusivité de la recherche. On pourrait imaginer l'utilisation de notions de sociolinguistique pour analyser des biais socio-économiques ou de la typologie pour améliorer les modèles multilingues (Hofmann et al (2024)).

*Intervenant

1https://en.wikipedia.org/wiki/Wikipedia:Who_writes_Wikipedia%3F