Biais liés aux ressources dans les systèmes à base d'apprentissage automatique

DJAMÉ SEDDAH

Inria Paris

djame.seddah@inria.fr

Ce panel de discussion a eu pour but de faire émerger des questionnements et de susciter des échanges entre participants aux journées jointes. Pour ce faire, nous avons utilisé l'outil https://speakup.info.

Dans une première phase, les participants ont rassemblé un grand nombre de questionnements, parmi lesquels :

- Comment identifier de nouveaux types de biais, en particulier ceux qui peuvent être des "menaces" pour les tâches considérées ?
- Pour les langues peu dotées ou les petits jeu de données, peut-on accepter de fermer les yeux sur les éventuels biais présents dans les données disponibles ou ne doit on pas sacrifier la préoccupation de neutralité ?
- Est-ce acceptable d'utiliser des données de faible qualité, ou biaisées, lorsque l'on est dans un contexte à faibles ressources ?
- Les biais ne sont ils pas nécessaires pour représenter le monde dans lequel nous sommes ?
- Doit on supprimer tous les biais ?
- Est-il possible de "quantifier"/définir correctement les biais pour séparer les données "biaisées" des données alors "non biaisées" ? Si oui, comment le définir ?
- Comment bien documenter des données d'apprentissage ?

- Est-ce que l'explicabilité / l'interprétabilité des modèles peut aider à contourner les effets néfastes des biais dans les données?
- Noms de professions et occupations féminins (directrice, créatrice etc). Doit-on les introduire artificiellement dans nos données ? De même façon pour toutes les langues ?
- Quels sont les biais associés aux ressources/modèles créés pour les langues peu dotées ?
- Qui décide de ce qui constitue un biais ?
- Question sur la temporalité des biais : est-ce qu'on peut détecter leurs évolution au cours du temps ?
- Comment faire pour contourner le biais de disponibilité de grandes quantités de données sur les tâches sur lesquelles nous travaillons ?
- Quels sont les biais induits par les données du web utilisées pour l'apprentissage des modèles de TAL ?
- Faut-il vraiment poser la question des biais pour des artefacts scientifiques uniquement destinés à l'étude par des professionnels ? (distinction entre applications des modèles de langues [LMs] et Recherche sur les LMs, cas de Delphi)
- Quel est le statut juridique des modèles de TAL qui reposent sur des données publiques ?
- Quelle est la différence entre un facteur et un biais ?
- Les biais ayant une vision sociétale négative évoluent dans le temps. Est-ce que ça suggère que les biais doivent être traités post-hoc ?
- Quelles sont les méthodes pour supprimer l'accroissement du biais par les modèles ?
- Les modèles de langues ne doivent ils pas refléter les biais de la société ?
- Comment évaluer la qualité des données d'apprentissage (ex : erreurs d'OCR, etc.) ?