Les biais « politiques « des très grands modèles de langue multilingues

Stage M2 proposé par François Yvon

ISIR / MLIA , Sorbonne Université

Entraînés sur des giga corpus de textes, les très grands modèles de langue (LLM) ont fait la preuve de leur capacité à engendrer des textes qui sont utiles dans bon nombre de situations : réponse à des questions, résumé et traduction automatiques, etc. Les énoncés ainsi produits sont loin d'être idéaux : au dela des erreurs factuelles dont ils sont truffés, se pose également la question des nombreuses formes de biais qu'ils expriment. Ces biais peuvent porter sur le genre, d'ethnicité, de nationalité, d'âge, etc mais également se manifester par une préférence pour certaines opinions politiques [Liang et al., 2021, Gallegos et al., 2023]. Au vu des préjudices que ces biais peuvent causer, leur identification et leur correction sont devenus des préoccupations majeures de la communauté scientifique (et des régulateurs), alors que ces modèles prennent une place croissante dans la palette d'outils d'accès à l'information.

Dans le cadre de ce stage, nous nous intéressons à une application particulière des LLM, la traduction automatique, en nous interrogeant sur leur capacité à préserver les opinions subjectives exprimées dans les textes sources et à les retraduire fidèlement dans les traductions cibles. Pour ce faire, il est possible de considérer deux types d'énoncés : d'une part, les énoncés subjectifs, qui expriment une opinion ou un point de vue, et dont on souhaite que la traduction respecte le point de vue ; d'autre part les énoncés objectifs, qui visent à exprimer des connaissances factuelles, et dont on souhaite que que la traduction respecte une forme de neutralité du point de vue.

Ce stage vise donc à étudier les biais des systèmes de traduction utilisant des LLM multilingues, en étendant les travaux conduits sur les biais de genre [Bordia and Bowman, 2019, Savoldi et al., 2021] à d'autres types de biais susceptibles d'entraîner une distortion du point de vue exprimé dans le texte source. Un type de biais particulièrement intéressant dans ce contexte est le biais »politique« [Doan and Gulla, 2022, Feng et al., 2023]. Parmi les travaux à accomplir pendant le stage :

— préparation d'un état de l'art sur la définition et l'identification du degré de subjectivité des textes

— identification de ressources disponibles pour les langues française et anglaise (jeux de données pour l'apprentissage et l'évaluation)
— développement de systèmes multilingues pour la mesure du degré de subjectivité des énoncés
— étude de la préservation du caractère subjectif des textes traduits par des modèles de traduction et par des grands modèles de langue multilingues.

Contexte du stage

Ce travail se déroulera idéalement de mars à juillet 2024. Il sera encadré par F. Yvon, chercheur à l'ISIR au sein de l'équipe MLIA, et gratifié selon les règles en vigueur à Sorbonne Université. Contact : yvon@isir.upmc.fr.

Références

Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi : 10.18653/v1/N19-3002. URL https://aclanthology.org/N19-3002.

Tu My Doan and Jon Atle Gulla. A survey on political viewpoints identification. *Online Social Networks and Media*, 30 :100208, 2022. ISSN 2468-6964. doi : https://doi.org/10.1016/j.osnem.2022.100208. URL https://www.sciencedirect.com/science/article/pii/S246869642200012X.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks : Tracking the trails of political biases leading to unfair nlp models. 05 2023. URL https://arxiv.org/pdf/2305.08283.pdf.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models : A survey. 09 2023. URL https://arxiv.org/pdf/2309.00770.pdf.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liang21a.html.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9 :845–874, 2021. doi : 10.1162/tacl_a_00401. URL https://aclanthology.org/2021.tacl-1.51.