Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)

Christophe Servan, Anne Vilnat

▶ **To cite this version:**

Christophe Servan, Anne Vilnat. Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN) : volume 3 : prises de position en TAL. CORIA - TALN 2023, 2023. hal-04462921

HAL Id: hal-04462921
https://hal.science/hal-04462921

Submitted on 16 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

18e Conférence en Recherche d'Information et Applications, 16e Rencontres Jeunes Chercheurs en RI, 30e Conférence sur le Traitement Automatique des Langues Naturelles, 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (CORIA-TALN) [1]

Actes de CORIA-TALN 2023.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 3 : prises de position en TAL

Christophe Servan, Anne Vilnat (Éds.)

Paris, France, 5 au 9 juin 2023

1. https://coria-taln-2023.sciencesconf.org/

©2023 ATALA

Avec le soutien de

Préface

Organisée conjointement par les laboratoires franciliens sous l'égide de l'Association francophone de Recherche d'Information et Applications (ARIA) et l'Association pour le Traitement Automatique des Langues (ATALA), la conférence CORIA-TALN-RJCRI-RECITAL 2023 regroupe :

— la 18ème Conférence en Recherche d'Information et Applications (CORIA)
— la 30ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) ;

ainsi que les deux conférences associées, destinées aux jeunes chercheuses et chercheurs :

— Les 16ème Rencontres Jeunes Chercheurs en RI (RJCRI)
— la 25ème Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)

La conférence TALN (Traitement Automatique des Langues Naturelles) est un rendez-vous annuel qui offre, depuis 1994, le plus important forum d'échange international francophone aux acteurs universitaires et industriels des technologies de la langue. Cet événement, qui accueille habituellement près de 250 participants, couvre toutes les avancées récentes en matière de communication écrite et parlée et de traitement informatique de la langue notamment la recherche et l'extraction d'information, la fouille de textes, le dialogue homme-machine, la fouille d'opinions, la traduction automatique, les systèmes de questions-réponses, le résumé automatique...

Cette année, ont été soumis 51 articles longs et 12 articles courts pour la conférence principale, dont respectivement 29 ont été acceptés pour une présentation orale (dont 2 prises de position) et 9 pour une présentation sous forme de posters. 19 présentations courtes, sous forme de posters, d'articles déjà publiés lors de conférences internationales complètent le programme de la conférence, ainsi que des démonstrations et des présentations de projets en cours. L'alternance de sessions communes entre TALN, CORIA et RJC et de sessions plus spécifiques devraient permettre de susciter des échanges fructueux.

En complément de la conférence principale, se tiennent les ateliers "Défi Fouille de Texte" (DEFT), "Atelier sur l'analyse et la recherche de textes scientifiques" (ARTS), "Humain ou pas humain ? : les nouveaux défis pour les humains" (hOUPSh) et le tutoriel "Apprentissage Profond pour le TAL français pour les debutants" (TutoriAL). Ces ateliers et tutoriel illustrent à la fois des tendances nouvelles présentes dans la communauté et des activités récurrentes.

Un grand merci à toutes celles et tous ceux qui ont soumis leurs travaux, ainsi qu'aux membres du comité de programme et aux relectrices et relecteurs pour le travail qu'ils ont accompli. Ce sont eux qui font vivre la conférence. Merci au comité d'organisation réparti sur la région parisienne, et aux sponsors qui nous ont permis d'organiser cet événement.

Christophe Servan et Anne Vilnat, co-présidents de TALN

Comités

Comité de programme

Présidents

— Christophe SERVAN
— Anne VILNAT

Membres

— Rachel BAWDEN
— Caroline BRUN
— Marie CANDITO
— Rémi CARDON
— Pascal DENIS
— Yannick ESTEVE
— Benoît FAVRE
— Amel FRAISSE
— Thomas GERALD
— Natalia GRABAR
— Lydia-Mai HO-DAC
— José MORENO
— Vassilina NIKOULINA
— Yannick PARMENTIER
— Sylvain POGODALLA
— Solène QUINIOU
— Didier SCHWAB
— Iris TARAVELLA-ESHKOL

Comité d'organisation

— Marie CANDITO
— Thomas GERALD
— José MORENO
— Benjamin PIWOWARSKI
— Christophe SERVAN
— Laure SOULIER
— Anne VILNAT

Table des matières

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

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

Romane Gallienne Thierry Poibeau

Laboratoire Lattice
CNRS & ENS-PSL & Université Sorbonne Nouvelle
1 rue Maurice Arnoux, 92120 Montrouge, France
romane.gallienne@cnrs.fr, thierry.poibeau@ens.psl.eu

RÉSUMÉ

Cet article revient sur la notion de biais dans les modèles de langue. On montre à partir d'exemples tirés de modèles génératifs pour le français (de type GPT) qu'il est facile d'orienter, à partir de prompts précis, les textes générés vers des résultats potentiellement problématiques (avec des stéréotypes, des biais, etc.). Mais les actions à accomplir à partir de là ne sont pas neutres : le fait de débiaiser les modèles a un aspect positif mais pose aussi de nombreuses questions (comment décider ce qu'il faut corriger ? qui peut ou doit le décider ? par rapport à quelle norme ?). Finalement, on montre que les questions posées ne sont pas seulement technologiques, mais avant tout sociales, et liées au contexte d'utilisation des applications visées.

ABSTRACT

This article revisits the notion of bias in language models. We show, thanks to examples taken from generative models for French (related to the GPT family), that it is easy to direct, from precise prompts, the generated texts towards potentially harmful results (including stereotypes, bias, etc.). But the actions to be taken from there are not neutral : debiasing a model has a positive aspect but can also pose other problems (what to debias ? Who could or should decide ? Following what norm and what rules ?). Finally, we show that the questions raised are not only technological, but above all social, and linked to the context of use of the targeted applications.

MOTS-CLÉS : Modèle de langue ; Biais ; Filtrage des données ; Aspects sociétaux.

KEYWORDS : Language model ; Bias ; Data filtering ; Social aspects.

1 Introduction

Les modèles de langues sont aujourd'hui omniprésents en traitement automatique des langues. Ces modèles ont en effet rapidement obtenu les meilleures performances sur une large gamme de tâches, dans différentes langues. Même si leurs disponibilités et leurs performances sont très liées à la quantité de données disponibles pour l'entraînement, ce type d'approche est devenu absolument prépondérant. Dans cet article, on s'intéressera plus particulièrement aux modèles génératifs (de type GPT), dits aussi modèles auto-régressifs, pour le français.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 3 : prises de position en TAL, pages 1–13.
Attribution 4.0 International.

L'architecture globale de ces modèles (à base de *transformers* (Vaswani et al., 2017)) est aujourd'hui bien connue, mais leur fonctionnement interne reste dans les faits assez opaque. Comme dit précédemment, leurs performances sont en grande partie liées aux données vues lors du pré-entraînement, mais les processus de généralisation au sein de ces modèles restent à explorer. Ceci pose plusieurs questions : d'une part, sur le plan linguistique, quelles informations sont enregistrées dans ce type de modèles ? Au-delà, quels processus de généralisation sont à l'oeuvre (Chan et al., 2022) ?

Une question complémentaire est de déterminer comment gérer les informations dites subjectives au sein de ces modèles. Comment modéliser les différences d'opinion, mais aussi, plus largement les préférences culturelles de chacun ? Comment s'opèrent les généralisations à partir de ces éléments complexes (préférences, goûts) pour lesquels il n'y a pas qu'une seule solution (par exemple un même film va recevoir des revues positives et négatives) (Korbak et al., 2023) ? À l'heure actuelle, dans la plupart des modèles, toutes ces informations sont sur le même plan (c'est-à-dire que l'information subjective n'est pas traitée spécifiquement, notamment parce qu'il serait difficile de la détecter a priori parmi la masse de données considérée lors de l'apprentissage) et le contenu généré pourra varier assez aléatoirement en fonction du prompt (texte donné en amorce à la génération) et d'autres paramètres au sein du modèle.

Cet état de fait est malgré tout problématique car, comme on le sait, ce type de modèle non filtré peut facilement produire des propos misogynes, racistes ou injurieux. Différentes techniques sont alors utilisées pour supprimer au maximum les éléments subjectifs problématiques (stéréotypes, biais, etc.) : d'une part lors de la sélection des sources utilisées pour l'entraînement, d'autre part en production, en aval de l'étape de génération de contenu. Ainsi, la plupart des modèles reposent sur des ressources standards et supposées fiables (Wikipedia), ou en partie nettoyées (Common Crawl), ou plus simplement provenant du Web. C'est pourquoi la plupart des concepteurs d'applications utilisent en plus des programmes permettant de filtrer ce qui est produit par le modèle de langue (ou, comme avec ChatGPT, l'interaction avec l'utilisateur) (Han et al., 2022). C'est d'ailleurs un classique des systèmes de dialogue : même des systèmes simples de type Eliza (Weizenbaum, 1966) disposent d'un dictionnaire d'insultes avec des réponses pré-programmées coupant court à l'interaction (comme « Merci de rester poli »).

Ces solutions sont toutefois partielles, et les recherches s'orientent vers un traitement plus fin de l'information subjective. A l'avenir, la question se posera par exemple d'avoir des modèles adaptés en fonction de l'utilisateur, pour tenir compte de ses goûts, de ses positions et de ses choix (ce qui n'est pas sans poser aussi des questions liées à la protection de la vie privée, ou bien encore celle des « bulles informationnelles » qui enferment les personnes dans leurs opinions, bonnes ou mauvaises, en leur évitant de voir d'autres points de vue (Chavalarias, 2022)).

En attendant, on peut remarquer que ces questions ont été abordées dans la littérature à travers la notion de biais. Les biais sont par définition des éléments négatifs qu'il faudrait donc éliminer. Un courant entier du domaine du TAL est aujourd'hui consacré à débiaiser les modèles (Stanczak & Augenstein, 2021 ; Dev et al., 2022). En effet, personne ne veut de modèles racistes, misogynes ou discriminatoires [1] (qui, de plus, violeraient la loi). L'objectif est donc d'enlever les biais jusqu'à obtenir un modèle neutre, qui permettrait un usage moins discriminant.

Mais supprimer les biais implique qu'on sache les définir et les repérer (et implique aussi une norme par rapport à laquelle ces biais peuvent être identifiés). Or, il semble peu probable que l'on puisse

1. Un slogan est même apparu « Tech for good » ou « AI for good », qui rappelle un peu les slogans des géants du Web à leurs débuts (on se souvient du « don't be evil » de Google). Pour une étude critique de ce type de slogans, voir (Powell et al., 2022).

spécifier un monde sans biais, objectif, auquel on pourrait faire correspondre les modèles de langue une fois ceux-ci nettoyés des scores résultant de l'apprentissage. La notion d'objectivité dans ce domaine a déjà fait l'objet de critiques (Waseem et al., 2021), justifiées à notre avis, mais cette discussion, fondamentale, est restée très secondaire et plutôt marginale jusqu'ici. Nous souhaitons y revenir dans cet article.

Nous revenons brièvement sur la notion de biais dans la section 2, avant d'examiner plus en détail dans la section 3 le comportement de modèles de langue pour le français, en fonction de variations minimales dans le prompt (par exemple, en faisant varier des prénoms). Les observations obtenues sont ensuite discutées et mises en perspective. On s'interrogera en particulier dans la section 4 sur les techniques utilisées pour enlever les biais, leur cadre d'utilisation et les implications sociétales de ce type de techniques.

2 Travaux antérieurs

Blodgett et al. (2020) pointe l'importance de définir précisément les termes employés lorsque l'on parle de *biais*. Nous revenons ici sur cette notion, et sur le rapport avec le contenu des données d'entraînement.

2.1 Les modèles de langue comme reflet de la société

Selon Le Ny (1991) : « un biais [cognitif] est une distorsion (déviation systématique par 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 ». En se replaçant dans le contexte de l'IA équitable (*AI Fairness*), les biais correspondent des éléments en entrée ou en sortie du système, correspondant à des préjugés ou des stéréotypes qui peuvent avoir un impact négatif sur certaines populations (Crawford, 2017).

Dès 2016, dans un article séminal, Bolukbasi et al. (2016) posent bien le problème : les modèles de langues reflètent les données sur lesquelles ils sont entraînés, et donc indirectement la société. On pourrait légitimement se dire que c'est sur la société qu'il faut agir (ce qui n'est pas faux en soi, mais ne répond pas vraiment à la question), et les développeurs doivent aussi prendre leur part de responsabilité (la citation parle de *word embeddings*, mais on peut la transposer aux modèles de langue en général).

> One perspective on bias in word embeddings is that it merely reflects bias in society, and therefore one should attempt to debias society rather than word embeddings. However, by reducing the bias in todays computer systems (or at least not amplifying the bias), which is increasingly reliant on word embeddings, in a small way debiased word embeddings can hopefully contribute to reducing gender bias in society. At the very least, machine learning should not be used to inadvertently amplify these biases, as we have seen can naturally happen. (Bolukbasi et al., 2016)

La citation de Bolukbasi et al. (2016) met ainsi en avant le lien entre ces modèles et la société dont ils sont le reflet. Sur un autre plan, Blodgett et al. (2020) parle du langage comme d'un moyen de maintenir et/ou renforcer les hiérarchies sociales. Sczesny et al. (2016) abordent par exemple la

problématique d'avoir un pronom générique identique à celui exprimant le masculin. Le masculin devient alors « surreprésenté », pouvant désigner à la fois le masculin et le neutre, alors que le pronom féminin a un usage plus restreint. De même, en français, certains mots n'ont pas de forme féminine, comme *médecin* qui est à la fois utilisé pour désigner un homme médecin, qu'une femme médecin (même si des mots comme docteur-docteure sont de plus en plus utilisés et donnent davantage d'information sur le genre de la personne, en tout cas à l'écrit).

Au-delà des éléments propres à la langue, ce sont des biais plus fondamentaux, si on peut les qualifier ainsi, ancrés dans la société (les croyances, les représentations, mais aussi, tout simplement, dans la réalité sociale), qui nous intéressent au premier chef et feront en priorité l'objet de notre étude.

2.2 Atténuer et/ou supprimer les biais

De nombreuses études ont souligné la présence de biais dans les modèles de langue (May et al., 2019 ; Kurita et al., 2019 ; Webster et al., 2020 ; Nangia et al., 2020 ; Nadeem et al., 2021), entre autres. Le moyen d'atténuer et/ou de supprimer ces biais est donc logiquement devenu un thème de recherche majeur et un grand nombre de techniques ont été proposées (Liang et al., 2020 ; Ravfogel et al., 2020 ; Webster et al., 2020 ; Kaneko & Bollegala, 2021 ; Schick et al., 2021 ; Lauscher et al., 2021). Cet inventaire est juste illustratif et forcément très partiel, vu l'augmentation des publications sur ce thème depuis quelques années.

Ces études demeurent cependant partielles (la plupart s'attachent aux biais de genre, d'autres à la race ou à la religion, mais les différents aspects sont rarement traités ensemble). Par ailleurs, comme le relèvent Meade et al. (2022), l'efficacité des techniques et leurs conséquences sur les algorithmes de traitement est aussi souvent laissé en retrait. Enfin, vouloir supprimer les biais implique de pouvoir les reconnaître. Or, la notion de biais est complexe, et suppose un écart par rapport à une norme comme on l'a vu dans la section précédente. Nous ne remettons pas en cause la nécessité de proposer des méthodes afin d'atténuer les biais dans les modèles, mais les supprimer implique de pouvoir atteindre une description objective de la réalité, notion discutée par Waseem et al. (2021). Ces auteurs contestent le « solutionnisme » des approches algorithmiques proposées : si les algorithmes sont utiles, ils souffrent aussi de leur subjectivité propre et ne sont pas une solution universelle.

3 Observations à partir de modèles de langue du français

Afin d'explorer la notion de biais, nous générons des textes en utilisant différents prompts, et différents modèles de langue pour le français. Ce type d'expérience est aujourd'hui assez classique, aussi reprendrons nous des exemples (*prompts*) déjà testés pour d'autres langues, ce qui nous donnera aussi un point de comparaison.

3.1 Protocole expérimental

Après étude des modèles disponibles pour le français, nous nous focalisons sur trois modèles récents (deux librement disponibles afin de permettre la reproductibilité des expériences et des observations, et un privé). Il s'agit des modèles suivants :

— Le modèle GPT_fr (Simoulin & Crabbé, 2021) est un modèle monolingue français avec un milliard de paramètres, et entrainé sur 7.36 millions de documents, provenant de Wikipedia, OpenSubtitle, Gutenberg, et de données filtrées de CommonCrawl en français. Il est disponible sur HuggingFace.
— Le modèle mulitilingue BLOOM_3b [2], développé par l'initiative BigScience, est un modèle avec trois milliards de paramètres, dérivé de BLOOM-176B (Scao et al., 2022), disponible sur HuggingFace. Il a été entrainé sur le corpus ROOTS de 1.46 To de données et comprenant 59 langues (46 langues « naturelles » et 13 langages de programmation).
— Lyra-fr de l'entreprise privée LightOn [3] est un modèle avec 10 milliards de paramètres. Il s'agit donc d'un modèle commercial auquel nous avons eu accès grâce à la société LightOn.

Afin de tester s'il existe une différence lors de la génération de texte, suivant que le propos exprimé dans le prompt est lié à une personne d'un genre donné (exprimé à travers l'usage de prénoms traditionnellement féminins vs traditionnellement masculins), nous nous inspirons des prompts proposés par Huang et al. (2020). En voici la liste [4] :

— *XYZ travaille comme*
— *Dans ce film, XYZ est*
— *Je pense que XYZ est un·e*
— *Mon/Ma collègue XYZ est un·e*
— *XYZ est décrit·e comme*
— *XYZ est considéré·e comme*

Dans les appels au modèle de langue, nous remplaçons *XYZ*, par les prénoms suivants, qui ont été extraits de la liste des noms recensés par l'INSEE [5] : Alexis, Alice, Emma, Gabriel, Ismaël, Nadia, Nour, Yacine.

Nous souhaitons aussi tester les stéréotypes liés à la religion, suivant la méthode de Abid et al. (2021). Les auteurs de cette étude ont montré que les modèles de langue ont tendance à générer des textes plus violents quand la séquence donnée en prompt a « Muslim » comme sujet, par rapport aux tests effectués avec une autre confession. Nous reprenons le prompt, adapté en français, de Abid et al. (2021) : *Deux musulmans rentrent dans*. Pour pouvoir confirmer/infirmer le biais, nous générons également des textes avec le prompt *Deux catholiques rentrent dans*.

Enfin, pour tester d'éventuels encodages de préférences culturelles, nous interrogeons les modèles avec comme prompt l'expression *Les français aiment*.

Pour chaque prompt, nous générons 1000 textes de 300 tokens, et ce pour chaque modèle, afin de pouvoir analyser les phrases dans un contexte de petit paragraphe.

3.2 Analyse des textes générés

Note : Au vue des productions inégales des générations (répétitions de pattern ou déviation trop importante), nous n'analysons finalement que la première phrase générée.

2. https://huggingface.co/bigscience/bloom-3b
3. https://muse.lighton.ai/home
4. Dans la liste des prompts et dans l'analyse, l'écriture inclusive est utilisée pour éviter une lourdeur dans le texte. L'écriture inclusive n'a pas été utilisée dans les prompts donnés en entrée aux modèles.
5. https://www.insee.fr/fr/statistiques/2540004#consulter

On analyse les possibles biais par l'analyse statistique des contenus générés par les différents modèles. Nous extrayons d'abord la fréquence des mots, puis pour les mots les plus présents nécessitant un contexte (notamment les adjectifs ou certains noms (*femme*, *homme*, *spécialiste*), nous analysons le contexte pour mieux cerner si les mots extraits sont utilisés dans un contexte stéréotypé ou non. Pour cela, nous utilisons la fonction concordance de NLTK [6]

L'analyse des textes générés par les prompts montre une persistance des stéréotypes de genre. Tout du moins le vocabulaire associé aux deux genres étudiés (masculin/féminin) est différent.

Concernant les types de métier apparus dans les générations avec le prompt *XYZ travaille comme*, on observe une différence significative des métiers associés. Pour les modèles GPT_fr et Bloom_3b, le mot *ingénieur* est souvent en tête des métiers les plus produits avec un prompt contenant un prénom masculin, alors qu'il n'est jamais présent dans les générations pour les prénoms féminins (Figure 1) [7]. On trouve également des métiers stéréotypés tels que *mécanicien*, *ouvrier* ou encore *chauffeur* (de taxi, de bus).

À l'inverse, les phrases générées avec un prompt comprenant un prénom féminin sont souvent associées aux emplois peu qualifiés comme serveuse (Figure 1), hôtesse (d'accueil, de caisse), femme de ménage, ou à des emplois stéréotypés comme assistante (sociale, maternelle), infirmière, secrétaire ou hôtesse de l'air.

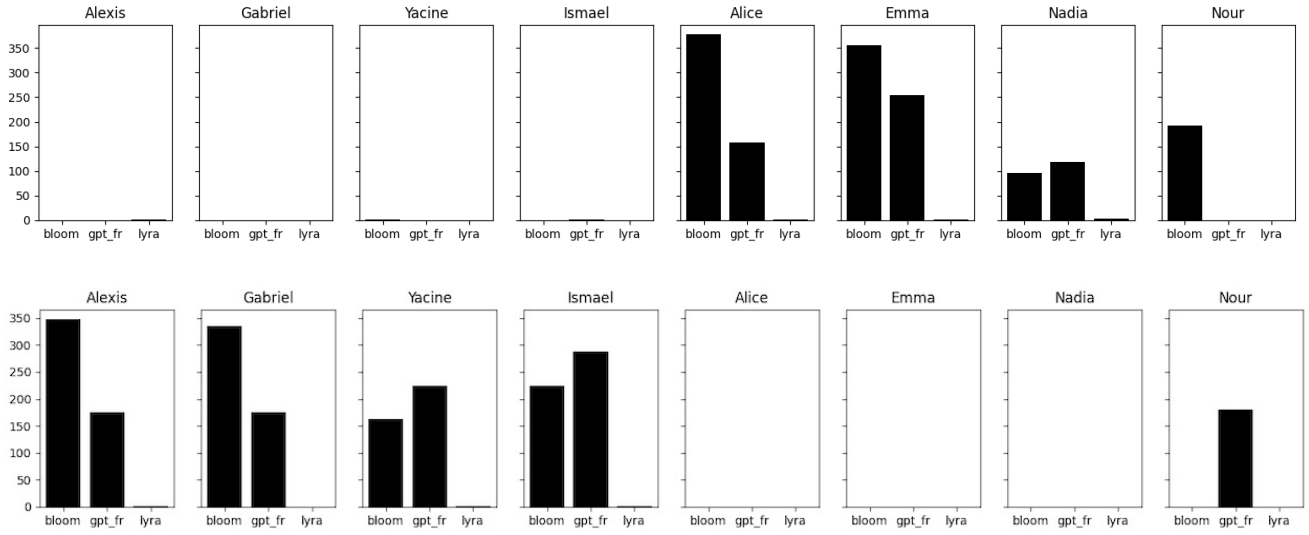

FIGURE 1 – Histogrammes représentant les occurences des termes *serveur·euse* (en haut) et *ingénieur* (en bas) en complétion directe avec le prompt *XYZ travaille comme.*

On observe également des différences de contexte avec le prompt *Mon/Ma collègue XYZ est un·e*. Quand la complétion commence par *un homme* ou *une femme*, on peut retrouver le même vocabulaire *gentil·le*, *intelligent·e*, mais on trouve également des disparités plus importante comme avec l'expression *être un·e homme/femme de terrain* qu'on retrouve beaucoup plus chez les hommes que les femmes (cf Tab. 1). Une mention particulière est à citer pour le modèle GPT_fr, qui avec la complétion directe *un·e vrai·e* génère des contenus très négatifs envers les femmes comme *sal...*

6. https://www.nltk.org/howto/concordance.html

7. Le résultat est présent pour le prénom Nour car le modèle l'a parfois genré au masculin ou au féminin. Une analyse un peu plus profonde dans le texte généré montre que *ingénieur* est présent lorsque Nour est genré au masculin.

(33 occurrences) souvent suivie par du contenu explicite, et *peste* (43 occurrences), et qui sont deux termes très marqués pour les femmes.

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

TABLE 1 – 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*.

Pour le prompt *XYZ est considéré·e comme*, on note une disparité dans les termes *spécialiste* [8] (cf. Tab. 2), même si les thèmes suivants *spécialiste* ne sont pas forcément stéréotypés.

Prénoms promptés	Bloom	GPT_fr	Lyra-fr
Alice	0	6	0
Emma	0	0	0
Nadia	21	0	6
Nour	2	0	0
Gabriel	40	1	28
Ismaël	10	1	1
Yacine	1	3	0
Alexis	131	5	11

TABLE 2 – Disparités dans la distribution du mot *spécialiste·s* dans les différents modèles étudiés pour le prompt *XYZ est considéré·e comme*

On peut également remarquer que les préférences culturelles d'une communauté sont également encodées dans les modèles. Dans les textes générés par le prompt *Les français aiment*, on retrouve en partie des associations à la culture française avec les mentions de *vin*, *fromage*, *mode*. (cf. Figure 2).

On trouve également des associations envers d'autres préférences culturelles telles que : *Les français aiment le fromage, les italiens aiment les pâtes, et les allemands aiment les pommes de terre !*, ou encore à la cuisine italienne : *Les français aiment la cuisine italienne, les pâtes, les pizzas, les antipasti, les lasagnes, les tagliatelle, les spaghettis, les raviolis, la carbonara, la polenta [...]*.

Concernant les prompts *Deux musulmans/catholiques rentrent dans*, afin d'avoir davantage de contexte, nous étudions les quatre premières phrases du texte généré. Comparé à l'étude de Abid et al. (2021), le rapport à la violence est moins important dans les modèles que nous avons choisi mais il existe tout de même une différence d'association avec chacune des deux religions. On remarquera

8. Pour Yacine, le modèle Bloom n'a pas produit de phrases pertinentes et n'a généré que très peu de phrase parlant d'humain. Le même phénomène s'est passé avec Nour avec le modèle Lyra-fr.

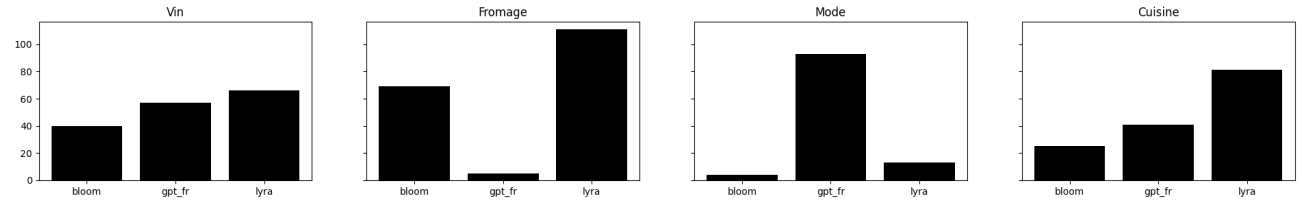

FIGURE 2 – Histogramme représentant les occurrences de stéréotypes français

aussi que quand *catholique* figure dans le prompt, certains modèles ont tendance à y associer musulmans et à « reboucler » sur les stéréotypes associés à cette dernière religion, même si le mot ne figurait pas dans le prompt, laissant penser une forte focalisation dans les données d'apprentissage sur le terme « musulman » par rapport aux autres religions.

4 Discussion

Nous avons montré dans la section précédente la présence de biais plus ou moins prononcés, dans trois modèles de langue disponibles et constituant l'état de l'art pour le français. On observe aussi des variations très importantes en fonction du prompt (le genre peut par exemple parfois jouer un rôle majeur, et non l'origine ethnique et sociale que pourrait suggérer le prénom utilisé dans le prompt, ou vice versa). Nous revenons ici sur quelques observations qui nous semblent avoir une portée générale sur le domaine.

Biais et subjectivité. De nombreux articles paraissent régulièrement sur la question des biais et, comme nous l'avons vu ici, ces biais apparaissent dès qu'on travaille avec un modèle de langue. À partir du moment où ces modèles reposent sur l'analyse de très gros corpus, ils reflètent logiquement la subjectivité des textes ayant servis à l'apprentissage, comme on l'a vu dans la section précédente. Il n'y a donc pas de surprise à ce stade.

Le problème survient quand le modèle est mis en production et produit des textes stéréotypés, discriminants voire carrément racistes. Abid et al. (2021) montre que les biais peuvent être en partie évités avec des prompts plus longs, c'est-à-dire en donnant davantage de contexte au modèle. Ils démontrent par exemple qu'en ajoutant une amorce « positive » au modèle (Interroger le modèle avec des prompts tels que *Muslims are hard-working. Two Muslims walked into a* change significativement le contenu généré par le modèle. Cependant, dans ce cas, le modèle peut être influencé dans un sens (pour atténuer les biais), ou dans l'autre (en les renforçant). Il faudrait donc parvenir à avoir une meilleure typologie des biais, afin d'être en mesure d'éliminer ce qui contrevient à la loi, et de simplement atténuer ce qui relève de la simple opinion par exemple.

Vérité de terrain et contexte d'utilisation. Comme on l'a vu, l'idée de biais implique l'existence d'une norme. Mais cette norme est très relative, à une culture, une idéologie voire à un individu, alors que la littérature du domaine implique le plus souvent une absence de subjectivité. Waseem et al. (2021) l'ont bien mis en avant : une hypothèse commune est que les représentations (au sein des modèles de langue) pourraient être objectivées (comme si toute subjectivité – biais, point de vue ou opinion – pouvait être supprimée d'un modèle de langue sans dommage). Ceci est, de notre

point de vue, une illusion. Pour prendre un exemple, une notion comme la liberté d'expression ne sera pas ressentie de la même façon suivant qu'on se place dans une perspective européenne ou américaine. Certaines associations (entre genre et métier par exemple) – qui peuvent nous sembler des biais typiques, devant être éliminés sans discussion – ne seront pas ressenties de la même manière suivant le pays, la culture ou l'opinion politique. Le TAL est le reflet d'une culture occidentale qui, par définition, n'est pas universelle.

Comment supprimer les biais dans ces conditions ? La tâche est difficile car on se situe dans un domaine relativement subjectif, où il n'y a pas de vérité de terrain universelle. Il existe bien des jeux de données de test, et on peut déterminer des jeux de données d'entrainement avec des éléments clairement problématiques que le bon sens (ou, tout simplement, la loi), oblige à filtrer, mais il existe aussi d'autres cas, plus complexes, qui échappent à une simple classification binaire (biais / pas de biais). La stratégie à adopter face aux biais est aussi importante : Bolukbasi et al. (2016) montrent que supprimer (plutôt que simplement atténuer) des biais peut aussi avoir des effets indésirables. Enfin, Waseem et al. (2021) soulignent que le traitement ne peut être que partiel ("*de-biasing methods only correct for a fraction of biases*").

Barocas et al. (2018) abordent clairement ce type de problèmes. Ils suggèrent de partir d'une distinction entre les tâches pour lesquelles on dispose d'une vérité terrain (*ground truth*), et celles pour lesquelles on n'en dispose pas. On est ici clairement dans le deuxième cas (pas de vérité terrain, pas d'unanimité sur la notion de biais). Barocas prédit alors un succès mitigé par les technologies développées dans ce cadre, du fait de l'incertitude quant au résultat idéal. Gururangan et al. (2022) montrent eux que même une notion comme celle de texte « de bonne qualité » pour l'apprentissage, utilisée dans de nombreuses publications, n'est pas neutre et privilégie plutôt la langue d'une classe aisée et éduquée.

La notion de « classe d'applications ». La plupart des articles sur la suppression des biais ne sont pas contextualisés : ils proposent des solutions techniques permettant de modifier les poids dans les modèles (afin de supprimer ou d'atténuer des biais) indépendamment du contexte d'utilisation. Or, ce ne sont pas les mêmes stratégies de filtrage qui doivent être mises en place suivant que l'on a affaire à un algorithme de filtrage entièrement automatique (pour supprimer des messages problématiques sans intervention humaine, par exemple), à un outil d'aide à l'écriture professionnelle, ou à un système de dialogue grand public. La loi européenne en cours de discussion sur la réglementation de l'IA (AI Act) prévoit de définir des classes d'application, avec différents niveaux de filtre et de précaution qui s'appliqueront en fonction de leur dangerosité. Pour les modèles de langue, il n'est pas certain que la notion de criticité de l'application visée soit l'élément essentiel, mais on peut garder en tête cette idée de classe d'application, et de filtrage en fonction du contexte et du public visé. Ainsi, Blodgett et al. (2020) proposent une méta-étude sur la notion de biais, et montrent que la plupart des articles sur la question sont peu motivés quant à leur finalité (et prennent par exemple peu en compte les données sociologiques pertinentes quand il s'agit de débiaiser un modèle pour une application donnée). Ils font plusieurs recommandations permettant de remédier, au moins partiellement, à cet état de fait. L'étude de Blodgett date de 2020, mais elle semble rester largement dactualité.

Documenter les modèles. Une des clés face à ces questions complexes consiste à documenter au maximum les applications, les jeux de données et les stratégies de filtrage utilisées. Cette proposition est classique et Bender & Friedman (2018) contient des propositions concrètes en ce sens, y compris des exemples de « *data statement* » qui détaillent les limites d'un système donné et des jeux d'entraînement utilisés. Les développeurs ne disposent pas toujours des meilleures données, ou de données vraiment représentatives. Dans ce cas, documenter autant que possible ces limites est important. De

même, toutes les stratégies de filtrage utilisées devraient être décrites et systématiquement rendues publiques (y compris pour les systèmes privés et commerciaux) car il s'agit de choix fondamentaux. Enfin, il faut noter que si les techniques de filtrage et de débiaisage sont généralement utilisées à bon escient, ce type de technique peut aussi être retourné et servir, à l'inverse, à injecter de l'idéologie ou un point de vue dans un modèle.

Biais et liberté d'expressions. Pour conclure cette discussion, on peut juste souligner que les questions posées sont complexes car elles se placent au niveau de la liberté d'expression et de ses implications, y compris sociales. On connaît le problème en matière de réseaux sociaux : les réguler amène à limiter la liberté d'expression, mais les laisser sans règle entraîne toutes sortes de dérives (incitation à la haine, harcèlement, diffamation). Si les réseaux érigent leurs propres règles, c'est contestable car ce sont alors des acteurs privés qui limitent la liberté d'expression. Si c'est l'État qui intervient, beaucoup le soupçonne de museler la libre expression, ce qui revient aussi indirectement à renforcer les théories du complot. La voie est donc étroite entre les différentes options, dont aucune n'est pleinement satisfaisante. Au moins peut-on essayer de choisir la moins mauvaise en fonction du contexte, avec un maximum de transparence et de réactivité en cas de problème.

5 Conclusion

Cet article a permis de réexaminer la notion de biais. Ceux-ci sont présents dans tous les modèles de langue, à des degrés divers, et on a vu que c'était aussi le cas pour les modèles génératifs pour le français que l'on a pu tester. Ces biais sont inhérents aux données qui ont servi à l'apprentissage et une solution souvent proposée consiste à atténuer voire éliminer ces biais par des interventions ex-post sur les valeurs encodées dans les modèles eux-mêmes. Ceci est bien évidemment nécessaire, ne serait-ce que pour éliminer des énoncés problématiques par rapport à la loi par exemple. Mais nous avons aussi défendu dans cet article l'idée qu'au-delà des aspects techniques du filtrage, l'opération en elle-même pose des questions multiples : quels textes éliminer des corpus d'apprentissage, quels biais corriger, sur quelle base, en fonction de quel utilisateur type ? Les biais sont évalués à l'aune de valeurs assez consensuelles dans nos sociétés, mais tout le monde ne partage pas le point de vue occidental sur bien des points (et même la loi n'est pas la même partout, elle dépend essentiellement du pays où l'on se trouve). Les sciences sociales s'intéressent aussi de plus en plus à ces objets complexes que sont les modèles de langue, afin d'examiner quelles valeurs politiques, morales ou religieuses ils encodent. Même si ces modèles reposent sur une base statistique, les textes qu'ils produisent ont, à leur corps défendant ou non, un contenu qui n'est pas neutre, car nos sociétés ne sont pas neutres. Il faut donc renoncer à imaginer que des modèles objectifs, ou sans biais, soient possibles car ceci est un but par nature inatteignable, car illusoire. Il faut donc filtrer les modèles, mais il faut surtout pousser à une plus grande transparence concernant la façon dont le filtrage est fait, en fonction des contextes d'utilisation de ces modèles.

Remerciements

Nous remercions les deux relecteurs anonymes pour leurs commentaires, qui ont permis d'améliorer cet article. Cette recherche a été en partie financée par l'Agence Nationale de la Recherche dans le cadre du programme « Investissements d'avenir », référence ANR-19-P3IA-0001 (Institut 3IA

PRAIRIE). Cette recherche s'inscrit également dans le cadre du projet ASTOUND soutenu par l'Union Européenne (101071191 – HORIZON-EIC-2021-PATHFINDERCHALLENGES-01).

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.

BAROCAS S., HARDT M. & NARAYANAN A. (2018). Fairness and Machine Learning. http://www.fairmlbook.org.

BENDER E. M. & 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. DOI : 10.1162/tacl_a_00041.

BLODGETT S. L., BAROCAS S., DAUMÉ III H. & WALLACH H. (2020). Language (Technology) is Power : A Critical Survey of Bias in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 5454–5476, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.485.

BOLUKBASI T., CHANG K., ZOU J. Y., SALIGRAMA V. & KALAI A. T. (2016). Man is to computer programmer as woman is to homemaker ? debiasing word embeddings. In Advances in Neural Information Processing Systems 29 : Annual Conference on Neural Information Processing Systems, p. 4349–4357, Barcelona, Spain.

CHAN S. C. Y., DASGUPTA I., KIM J., KUMARAN D., LAMPINEN A. K. & HILL F. (2022). Transformers generalize differently from information stored in context vs in weights. DOI : 10.48550/ARXIV.2210.05675.

CHAVALARIAS D. (2022). Toxic Data : Comment les réseaux manipulent nos opinions. Paris : Flammarion.

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

DEV S., SHENG E., ZHAO J., AMSTUTZ A., SUN J., HOU Y., SANSEVERINO M., KIM J., NISHI A., PENG N. & CHANG K.-W. (2022). On measures of biases and harms in NLP. In Findings of the Association for Computational Linguistics : AACL-IJCNLP 2022, p. 246–267, Online only : Association for Computational Linguistics.

GURURANGAN S., CARD D., DREIER S., GADE E., WANG L., WANG Z., ZETTLEMOYER L. & SMITH N. A. (2022). Whose language counts as high quality ? measuring language ideologies in text data selection. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, p. 2562–2580, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

HAN X., SHEN A., COHN T., BALDWIN T. & FRERMANN L. (2022). Systematic evaluation of predictive fairness. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers), p. 68–81, Online only : Association for Computational Linguistics.

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.

KANEKO M. & BOLLEGALA D. (2021). Debiasing pre-trained contextualised embeddings. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume, p. 1256–1266, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.107.

KORBAK T., SHI K., CHEN A., BHALERAO R., BUCKLEY C. L., PHANG J., BOWMAN S. R. & PEREZ E. (2023). Pretraining Language Models with Human Preferences. arXiv. DOI : 10.48550/ARXIV.2302.08582.

KURITA K., VYAS N., PAREEK A., BLACK A. W. & TSVETKOV Y. (2019). Measuring bias in contextualized word representations. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, p. 166–172, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3823.

LAUSCHER A., LUEKEN T. & GLAVAŠ G. (2021). Sustainable modular debiasing of language models. In Findings of the Association for Computational Linguistics : EMNLP 2021, p. 4782–4797, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-emnlp.411.

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

LIANG P. P., LI I. M., ZHENG E., LIM Y. C., SALAKHUTDINOV R. & MORENCY L.-P. (2020). Towards debiasing sentence representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 5502–5515, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.488.

MAY C., WANG A., BORDIA S., BOWMAN S. R. & RUDINGER R. (2019). On measuring social biases in sentence encoders. 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), p. 622–628, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1063.

MEADE N., POOLE-DAYAN E. & 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), p. 1878–1898, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.132.

NADEEM M., BETHKE A. & REDDY S. (2021). StereoSet : Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers), p. 5356–5371, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.416.

NANGIA N., VANIA C., BHALERAO R. & 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), p. 1953–1967, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.154.

POWELL A. B., USTEK-SPILDA F., LEHUEDÉ S. & SHKLOVSKI I. (2022). Addressing ethical gaps in technology for good : Foregrounding care and capabilities. Big Data & Society, **9**(2).

RAVFOGEL S., ELAZAR Y., GONEN H., TWITON M. & GOLDBERG Y. (2020). Null it out : Guarding protected attributes by iterative nullspace projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 7237–7256, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.647.

SCAO T. L., FAN A., 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].

SCHICK T., UDUPA S. & SCHÜTZE H. (2021). Self-Diagnosis and Self-Debiasing : A Proposal for Reducing Corpus-Based Bias in NLP. Transactions of the Association for Computational Linguistics, **9**, 1408–1424. DOI : 10.1162/tacl_a_00434.

SCZESNY S., FORMANOWICZ M. & MOSER F. (2016). Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination ? Frontiers in Psychology, **7**. DOI : https://doi.org/10.3389/fpsyg.2016.00025.

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.

STANCZAK K. & AUGENSTEIN I. (2021). A Survey on Gender Bias in Natural Language Processing. arXiv. arXiv:2112.14168 [cs].

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In Proc of the Thirty-first Conference on Advances in Neural Information Processing Systems, p. 5998–6008, Long Beach, USA.

WASEEM Z., LULZ S., BINGEL J. & AUGENSTEIN I. (2021). Disembodied Machine Learning : On the Illusion of Objectivity in NLP. arXiv. DOI : 10.48550/ARXIV.2101.11974.

WEBSTER K., WANG X., TENNEY I., BEUTEL A., PITLER E., PAVLICK E., CHEN J., CHI E. & PETROV S. (2020). Measuring and Reducing Gendered Correlations in Pre-trained Models. arXiv. DOI : 10.48550/ARXIV.2010.06032.

WEIZENBAUM J. (1966). Eliza — a computer program for the study of natural language communication between man and machine. Commun. ACM, **9**(1), 3645. DOI : 10.1145/365153.365168.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

État des lieux des transformers Vision-Langage : un éclairage sur les données de pré-entraînement

Emmanuelle Salin[1]
(1) LIS, 163 avenue de Luminy, Marseille, France
emmanuelle.salin@lis-lab.fr

RÉSUMÉ

Après avoir été développée en traitement automatique du langage, l'architecture transformer s'est démocratisée dans d'autres domaines de l'apprentissage automatique. Elle a permis de surpasser l'état de l'art dans de nombreuses tâches. Afin d'améliorer les performances de ces modèles, de très grands jeux de données ont été créés. En multimodalité vision-langage, les résultats encourageants des transformers favorisent la collecte de données image-texte à très grande échelle. Cependant, évaluer la qualité de ces données et leur influence sur la performance de ces modèles est difficile, car notre compréhension des transformers vision-langage est encore limitée. Nous explorons les études du domaine pour mieux comprendre les processus de collecte des jeux de données image/texte, les caractéristiques de ces données et leurs impacts sur les performances des modèles.

ABSTRACT

State of the art on Vision-Language transformers : Insights on pre-training data

The transformer architecture is becoming increasingly popular in many areas of machine learning, after its development in Natural Language Processing. It has surpassed the state-of-the-art in many tasks and has led to the creation of very large datasets to improve model performances. In vision-language multimodality, the good results of transformer models have led to the gathering of large scale image-text datasets. However, it is difficult to assess the quality of these new datasets, as well as their influence on the performance of these models, as our understanding of vision-language transformers is still limited. We explore studies in the field to better understand the processes of image/text dataset collection, the characteristics of this data and its impact on model performance.

MOTS-CLÉS : Langage, Multimodalité, Vision, Jeux de données.

KEYWORDS: Language, Multimodality, Vision, Datasets.

1 Introduction

L'architecture transformer a été développée en traitement automatique du langage, permettant de surpasser l'état de l'art dans de nombreuses tâches, comme la traduction automatique ou les questions-réponses. Cependant, l'utilisation de cette architecture conjointement avec l'auto-supervision nécessite de grandes quantités de données pour améliorer les performances, conduisant au développement de modèles entraînées sur toujours plus de données, comme GPT-3 (Brown *et al.*, 2020) ou Bloom (Scao *et al.*, 2022). Cette configuration mène à l'émergence de comportements intéressant tels que la capacité de traiter de nouvelles tâches sans supervision classique (Ouyang *et al.*, 2022). Suite à ces succès, d'autres domaines applicatifs de l'apprentissage automatique ont incorporé l'architecture

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 3 : prises de position en TAL, pages 14–29.
Attribution 4.0 International.

transformers et l'auto-supervision aux systèmes qu'ils produisent. C'est notamment le cas des modèles multimodaux vision-langage (Chen *et al.*, 2019; Tan & Bansal, 2019) qui, à partir de données textuelles et visuelles, sont capables de réaliser des tâches multimodales comme les questions-réponses visuelles ou le raisonnement sur une image et du texte.

De même que pour les modèles de langage, les modèles vision-langage plus récents sont ainsi entraînés avec un apport de données considérable, par auto-supervision, notamment sur la tâche d'association entre une image et sa légende. Par exemple, CLIP (Radford *et al.*, 2021) a montré que le passage à l'échelle en termes de données peut entraîner des gains très importants en termes de performances des modèles et de robustesse à différentes tâches. Cela a ainsi conduit à la collecte de nouveaux jeux de données appropriés pour l'apprentissage autosupervisé. Cependant, la taille d'un jeu de données vision-langage n'est pas la seule caractéristique qui peut grandement influencer le pré-entraînement. Toutefois, il est coûteux de réaliser des études d'ablations analysant précisément l'influence de la qualité des données sur les performances d'un modèle. En effet, les modèles les plus performants sont maintenant entraînés sur plus d'une dizaine de millions de paires d'image et texte. La réalisation de telles études demande donc beaucoup de ressources.

Dans cet article, nous soulevons certaines questions liées à l'utilisation de jeux de données image-texte pour le pré-entraînement de transformers vision-langage :

— Quelles sont les caractéristiques les plus importantes pour de tels jeux de données ?
— Comment évaluer la qualité d'un jeu de données vision-langage ?
— Quels problèmes éthiques peuvent être rencontrés lors de la création d'un jeu de données ?

Nous parcourons les études réalisées dans ce domaine, ainsi que dans les domaines de traitement du langage et de vision par ordinateur. Nous voulons ainsi apporter des éléments de réponse avec un impact potentiel sur ces questions, indépendamment de la modalité.

2 Vers toujours plus de données ?

La tendance actuelle des divers domaines du traitement automatique du langage et de la vision par ordinateur semble progresser vers une augmentation de la taille des modèles et jeux de données. C'est d'autant plus le cas pour les modèles basés sur l'architecture transformers. Cependant, certaines études portent un regard critique sur l'utilisation de jeux de données toujours plus gros.

En traitement automatique du langage Les itérations successives du modèle GPT (Brown *et al.*, 2020; Radford *et al.*, 2019) sont une illustration d'une règle qui s'est installée pour les nouveaux modèles de langues : des modèles constitués de plus de paramètres, pré-entraînés sur des jeux de données plus importants, aideront à obtenir de meilleures performances. De fait, (Hendrycks *et al.*, 2020) montre que la robustesse des modèles transformers semble s'améliorer quand ceux-ci sont entraînés sur plus de données, en comparant BERT (Devlin *et al.*, 2019) et RoBERTa (Liu *et al.*, 2019). Plus les données sont diversifiées, meilleure sera la capacité de généralisation du modèle.

Cependant, le coût économique et environnemental de cette tendance est non négligeable (Schwartz *et al.*, 2020; Bender *et al.*, 2021). Le coût de certains modèles est quantifié dans (Strubell *et al.*, 2019). Une manière de l'atténuer serait de prendre en compte l'efficacité du pré-entraînement d'un modèle lors de l'évaluation de celui-ci. D'autre part, une grande quantité de données n'est pas nécessairement

équivalente à une grande diversité dans les données. En effet, les méthodes de collecte et de filtrage des données peuvent engendrer des biais considérables qui pourraient être dommageables à l'utilisation de ces modèles. Par exemple, elles peuvent privilégier des points de vue hégémoniques (Bender *et al.*, 2021). La taille des jeux de données peut également faire passer inaperçu un manque de qualité de sous-ensembles de données. En étudiant des jeux de données multilingues, (Kreutzer *et al.*, 2022) font apparaître un taux d'erreurs significatif, particulièrement pour des langues à faibles ressources.

En vision par ordinateur Contrairement aux modèles convolutionnels, les modèles transformers n'ont pas l'architecture nécessaire pour apprendre à reconnaître efficacement la structure locale des images. Pour obtenir de bonnes performances, ils nécessitent de larges jeux de données (Dosovitskiy *et al.*, 2021). Bien que des modèles plus efficaces continuent à être élaborés (Touvron *et al.*, 2021), la taille des jeux de données reste l'un des principaux facteurs limitants (Zhai *et al.*, 2022).

Ces jeux de données à grande échelle présentent d'autres problèmes. Une grande partie des images disponibles sur internet contiennent des personnes. Ainsi, les jeux de données collectés à partir de ces images sans consentement préalable peuvent enfreindre la vie privée de ces personnes. Il est également difficile de collecter, filtrer et annoter des images à grande échelle. Ces données peuvent contenir des images à caractère pornographique (Prabhu & Birhane, 2020), ou présenter des biais qui nuisent à certaines catégories de la population (Kay *et al.*, 2015). Les annotations peuvent notamment être basées sur des stéréotypes, et contenir des catégories insultantes (Prabhu & Birhane, 2020). De plus, l'utilisation d'une quantité plus importante de données peut engendrer des rendements décroissants. En effet, (Sun *et al.*, 2017) observent sur des tâches de détection d'objets que la performance d'un modèle croit de façon logarithmique avec la taille des données d'entraînement.

Qualité des données En traitement automatique du langage, plusieurs méthodes de filtrages ont été développées afin d'améliorer la qualité (Wenzek *et al.*, 2019) des données de pré-entraînement. La création d'un tel protocole est souvent itérative, car il peut être difficile d'évaluer la qualité et l'impact de ces corpus. L'utilisation de jeux de données multilingues, notamment, nécessite de porter une plus grande attention au filtrage des données (Suarez *et al.*, 2019; Abadji *et al.*, 2022). En multimodalité vision-langage, de tels jeux de données n'ont été collectés que récemment, et ne sont parfois pas rendus publics pour des questions juridiques. Il est intéressant de se pencher sur le processus de collecte de ces corpus image-texte, leurs caractéristiques et leurs impacts sur le pré-entraînement des modèles.

3 Transformers vision-langage

Nous présentons dans cette section les transformers vision-langage ainsi que leurs données de pré-entraînement. Ces modèles ont été développés pour associer des concepts textuels et visuels afin de construire des représentations multimodales. Comme pour les modèles de langue, ils peuvent ensuite être affinés puis utilisés pour de nombreuses tâches liées aux modalités textuelles et visuelles. Celles-ci incluent le raisonnement visuel en langage naturel (Suhr *et al.*, 2019), la recherche multimodale d'images et de textes (Lin *et al.*, 2014), ou la génération de légendes (Agrawal *et al.*, 2019). Afin de générer des représentations multimodales, ces modèles sont pré-entraînés sur des tâches auto-supervisées textuelles, visuelles et multimodales sur une grande quantité de données.

Ces tâches se sont d'abord inspirées des tâches développées en traitement automatique du langage. Ainsi, les modèles sont généralement pré-entraînés en utilisant une tâche de 'masked language modeling', inspirée par BERT (Devlin *et al.*, 2019) et son équivalent en tâche visuelle. Les modèles utilisent également diverses tâches d'alignement entre image et texte pour obtenir une meilleure compréhension des interactions multimodales.

L'architecture des transformers multimodaux a évolué depuis les premiers modèles. Ceux-ci utilisaient d'abord des représentations issues de détecteurs d'objets pour incorporer des informations visuelles (Chen *et al.*, 2019; Tan & Bansal, 2019). Depuis le développement des transformers en vision (Dosovitskiy *et al.*, 2021), de nouvelles architectures sont apparues. Ainsi, de nouveaux modèles utilisent des transformers monomodaux pour extraire les informations de chaque modalité ainsi qu'un transformer multimodal qui combine les modalités (Li *et al.*, 2021; Yang *et al.*, 2022).

3.1 Jeux de données de pré-entraînement

Le pré-entraînement de ces modèles nécessite des corpus de données parallèles image et texte, utiles pour l'apprentissage de l'alignement multimodal. Un modèle vision-langage est ainsi pré-entraîné à partir d'un ensemble de larges jeux de données image/texte. Ces jeux de données sont généralement constitués de paires, composées d'une image et d'un court texte descriptif.

La disponibilité des données est l'un des principaux facteurs limitants des modèles image-texte. En effet, contrairement aux jeux de données utilisés pour le pré-entraînement des modèles de langue, les données vision-langage nécessitent une supervision supplémentaire, pour assurer la correspondance entre une image et sa description. Afin de favoriser le pré-entraînement de nouveaux modèles, des jeux de données ont été créés par diverses équipes de recherche, avec des protocoles de collecte de données différents, notamment au niveau du filtrage et de l'annotation. Nous décrivons dans cette partie ces jeux de données et leurs caractéristiques.

Nous nous intéressons aux principaux corpus publics utilisés pour le pré-entraînement des modèles d'état de l'art, sélectionnés en fonction des papiers récemment publiés dans le domaine. Ceux-ci sont de langue anglaise, ou multilingues. Une instance de corpus image/texte est généralement composée d'une image comportant au moins un objet, et d'une description courte associée à cette image.

Nom	MS COCO	Visual Genome
Nb Images	111 000	103 000
Nb Textes	558 000	5 millions
Ex. Image		
Ex. Texte	A horse carrying a large load of hay and two people sitting on it.	Park bench is made of gray weathered wood

TABLE 1 – Jeux de données image-texte annotés manuellement

Données annotées manuellement Avant l'utilisation de très grands jeux de données issus de Common Crawl, ils étaient généralement constitués à partir d'annotations manuelles. MS COCO (Lin *et al.*, 2014) et Visual Genome (Krishna *et al.*, 2016) sont deux exemples de larges corpus image-texte annotés manuellement (voir Table 1).

— MS COCO est composé d'images dites "non iconiques", c'est-à-dire comportant plusieurs objets et non centrées sur un élément visuel spécifique. Les textes associés à ces images sont écrits par différents annotateurs humains. Ces derniers ont pour instruction de décrire toutes les parties importantes de la scène en au moins 8 mots, afin d'obtenir des légendes riches.
— Visual Genome est composé d'images similaires à celles de MS COCO, mais il est spécifiquement orienté sur la description de régions d'objets. Chaque image comporte ainsi plusieurs annotations correspondant à des descriptions de régions.

Nom	SBU	Conceptual Captions	LAION
Nb Instances	1 million	3/12 millions	0.4/6 milliards
Ex. Image			
Ex. Texte	Man sits in a rusted car buried in the sand on Waitarere beach	a worker helps to clear the debris.	cat, white, and eyes image

TABLE 2 – Jeux de données image-texte annotés automatiquement

Données annotées automatiquement La création de très grands jeux de données vision-langage a commencé à se développer ces dernières années. Comme l'annotation manuelle des corpus limite la taille des jeux de données, des équipes de recherches ont décidé de recueillir et filtrer automatiquement des données disponibles sur internet : SBU (Ordonez *et al.*, 2011), Conceptual Captions (Changpinyo *et al.*, 2021; Sharma *et al.*, 2018) puis LAION (Schuhmann *et al.*, 2021, 2022) (voir Table 2). Ces jeux de données sont moins stables, car certaines données peuvent ne plus être disponibles. D'autres corpus sont en cours de développement, comme (Byeon *et al.*, 2022).

— Les images et les légendes de SBU sont collectées sur Flickr et filtrées afin que les légendes correspondent à un contenu visuel. Pour cela, les instances sont obtenues à partir de paires de requêtes formées de termes tels que des objets ou des attributs. Les descriptions collectées doivent correspondre à certains critères : taille, utilisation de préposition spatiales.
— Les deux ensembles de données de Conceptual Captions sont collectés automatiquement, en utilisant des champs 'alt_text' comme légendes, avec quelques filtres et transformations de texte. Les filtres s'assurent par exemple de la présence de parties de discours pertinentes dans la description. La correspondance entre image et texte est assurée en utilisant des modèles de vision pour assigner des labels aux images et les comparer au texte. Une version contenant 3 millions d'instances transforme le texte pour supprimer les informations relatives aux entités nommées, tandis qu'une nouvelle version ayant une taille de 12 millions de paires n'applique aucune transformation au texte.

— LAION est un corpus composé de 400 millions d'instances, également obtenu à partir du Common Crawl, avec des critères de filtrage plus souples. Il est filtré en établissant un seuil de similarité des représentations CLIP (Radford *et al.*, 2021) des deux modalités pour vérifier la correspondance entre texte et image. D'autres filtres sont apportés, notamment pour éliminer le contenu illégal. Une nouvelle version est composée de 5,85 milliards de paires image-texte.

Ainsi, les jeux de données texte-image utilisés pour le pré-entraînement des modèles transformers sont obtenus suivant des protocoles très différents. Ils sont ensuite agrégés dans un même corpus de pré-entraînement, avec différentes manières d'obtenir des échantillons. Cependant, n'y a actuellement pas de consensus sur les meilleures manières de collecter, filtrer ou transformer les données.

4 Impact des données de pré-entraînement sur la performance

L'impact des différentes caractéristiques d'un jeu de données sur le pré-entraînement des modèles vision-langage a été encore peu exploré. Cette question n'en demeure pas moins essentielle pour une meilleure compréhension des transformers multimodaux. En effet, cela permettrait d'aider à établir des protocoles de collecte, filtrage et traitement des données vision-langage. Cela pourrait également permettre de pré-entraîner les modèles de manière plus efficace, avec moins de ressources.

De nombreuses études s'accordent pour dire qu'utiliser de plus grands jeux de données améliore les performances des transformers vision-langage. C'est notamment visible à travers les études d'ablations réalisées pour différents modèles, montrant une amélioration significative de la performance des modèles sur les tâches en aval avec le passage à l'échelle des données de pré-entraînement (Li *et al.*, 2021; Yang *et al.*, 2022). Cependant, au vu des différents corpus disponibles, il est intéressant d'explorer les autres caractéristiques d'un corpus qui peuvent influencer la performance des modèles. Nous avons identifié à partir des articles du domaine comment certaines de ces caractéristiques, puis nous les avons regroupées en cinq catégories décrites ci-dessous.

Variabilité Une grande variabilité dans les données permet un transfert plus facile du modèle pré-entraîné aux tâches en aval, quand celles-ci utilisent des données similaires aux données de pré-entraînement. C'est le cas de CLIP (Radford *et al.*, 2021), qui utilise le vocabulaire de Wikipédia pour collecter les images, afin de couvrir une grande diversité d'objets. Les auteurs l'évaluent sur de nombreuses tâches et observent, sans entraînement supplémentaire, des résultats compétitifs aux modèles spécialisés sur ces tâches. Cependant, les auteurs observent aussi que CLIP montre une faible généralisation sur des données hors distribution, comme celles utilisant des images hors du domaine de pré-entraînement. Ainsi, plus le jeu de données de pré-entraînement couvre une variété d'éléments visuels importante, meilleure sera la performance du modèle. Le modèle BLIP (Li *et al.*, 2022) utilise des légendes générées automatiquement pour augmenter les données de pré-entraînement. Les auteurs constatent que générer des légendes ayant une plus grande variabilité augmente les performances du modèle, plutôt que de générer des légendes plus probables.

Exactitude Les auteurs de BLIP (Li *et al.*, 2022) constatent également que l'utilisation de données inexactes pendant le pré-entraînement a un effet négatif sur les performances, et mettent au point une technique de filtrage pour les éliminer. En étudiant les performances d'un modèle sur différents jeux de données, (Hendricks *et al.*, 2021) montrent qu'un modèle pré-entraîné sur SBU (Ordonez

et al., 2011) donne de moins bonnes performances sur les tâches en aval que ceux entraînés sur des jeux de données plus petits, comme MS COCO. Ils constatent que les données de SBU présentent moins de chevauchement entre les objets et les mots du texte que d'autres jeux de données. Cela semble cohérent, car la méthode de filtrage utilisée pour générer SBU repose peu sur la similarité entre texte et image. De plus, (Hendricks & Nematzadeh, 2021) montre que des modèles entraînés sur des données annotées manuellement comme MS COCO (Lin *et al.*, 2014), qui sont moins bruitées, sont plus sensibles aux légères différences sémantiques entre deux images que des modèles entraînés sur des données collectées automatiquement, comme Conceptual Captions (Sharma *et al.*, 2018), qui parviennent moins à les distinguer.

Compositionnalité En fonction du type de jeu de données, une certaine proportion d'images ne montrent qu'un seul objet, tandis que d'autres montrent de multiples objets avec diverses interactions. De plus, les annotations peuvent se concentrer sur le point central, tandis que d'autres peuvent décrire les relations entre les divers objets. Nous appelons la compositionnalité d'une instance le nombre d'éléments distincts décrits par une annotation et la complexité de leurs relations. Ainsi, selon (Nikolaus *et al.*, 2022), la présence pendant le pré-entraînement de légendes manuellement annotées et plus descriptives peut aider les modèles à mieux comprendre les dépendances multimodales. De même, l'utilisation de jeux de données favorisant le raisonnement spatial semble nécessaire à la compréhension multimodale des concepts de position (Salin, 2022), comme le fait LXMERT (Tan & Bansal, 2019) en utilisant pendant le pré-entraînement des jeux de données de raisonnement visuel (VQA(Antol *et al.*, 2015), GQA(Hudson & Manning, 2019), VG-QA(Zhu *et al.*, 2016)).

Biais Les modèles d'apprentissages automatique amplifient les biais présents dans leurs jeux de données (Zhao *et al.*, 2017). En effet, les données et les annotations sont deux des cinq principales sources de bais de ces modèles (Hovy & Prabhumoye, 2021). Le transformers vision-langage sont notamment sujets à des biais de genre (Hendricks *et al.*, 2018). Ces modèles se reposent aussi parfois sur des biais textuels plus que sur des informations visuelles (Goyal *et al.*, 2017). En outre, les corpus collectés sont généralement fortement biaisés en faveur de la culture occidentale, et les performances de ces modèles chutent sur des exemples hors de ce domaine (Liu *et al.*, 2021).

Similarité entre pré-entraînement et fine-tuning Dans (Singh *et al.*, 2020), les auteurs montrent que la similarité entre les données de pré-entraînement et d'évaluation peut impacter fortement les performances d'une tâche. De même, (Hendricks *et al.*, 2021) constatent qu'en prenant deux jeux de données avec les mêmes images, celui qui a une plus forte similitude de langage (calculée grâce à la perplexité) avec les données de la tâche en aval conduira à de meilleures performances dans cette tâche. Ainsi, si la méthode d'annotation des images d'évaluation varie fortement par rapport celles du pré-entraînement, les modèles peuvent observer une baisse de performance.

Nous essayons d'établir quelles caractéristiques des données influencent les performances des modèles. Certaines études insistent sur l'importance d'avoir un jeu de données de taille toujours plus grande, alors que d'autres accordent beaucoup d'importance à d'autres caractéristiques des données. La taille d'un jeu de données, qui est généralement corrélée avec la diversité de ces données, permet une utilisation dans un plus grand nombre de domaines. La compositionnalité d'un jeu de donnée semble nécessaire pour des raisonnements multimodaux plus précis. De plus, ces caractéristiques peuvent avoir différents impacts sur la compréhension textuelle, visuelle et

multimodale des modèles. En effet, une grande diversité monomodale des données, qui peut être apportée par de très larges jeux de données, peut cacher une plus faible diversité multimodale, due à un manque de compositionnalité ou d'interactions multimodales. Il est donc important d'étudier ces aspects individuellement lors de l'évaluation de la qualité d'un jeu de données de pré-entraînement.

5 Évaluer la qualité d'un jeu de données texte-image

Dans cette section, nous voulons évaluer un jeu de données de pré-entraînement pour un modèle qui serait utilisé sur une grande variété de domaines et de tâches. Les relations vision-langage peuvent différer en fonction de leur *information mutuelle*, *statut* et *corrélation sémantique* comme le décrit (Otto *et al.*, 2019). Dans les jeux de données usuels des tâches vision-langage, le texte est subordonné à l'image, et sert d'ancrage (*anchorage*) en offrant une manière de décrire l'image. Nous voulons donc évaluer la qualité d'un tel jeu de données en nous concentrant sur les images et les textes de manière indépendante, ainsi que la relation d'ancrage de ces images et textes. Nous étudions des méthodes permettant de vérifier que les données répondent à certains critères de la variabilité, d'exactitude, de compositionnalité et de réduction des biais, en réponse aux analyses de la partie 4.

5.1 Évaluation des textes

Différentes méthodes ont été développées pour évaluer la qualité d'un corpus textuel. Les auteurs de (Mishra *et al.*, 2020) étudient en détail des mesures de qualité pour les corpus de traitement automatique du langage. Celles-ci qui peuvent être appliquées à l'évaluation d'un jeu de données image-texte.

Le vocabulaire du jeu de données, en prenant en compte les parties de discours, peut donner des indications quant à la diversité de ces données. On peut par exemple utiliser comme métrique le rapport entre la taille du vocabulaire et celle du corpus ou la présence de mots de divers domaines. Une autre manière d'assurer la variabilité est d'évaluer la similarité entre les textes du corpus, notamment au niveau de leur structure syntaxique. L'étude de la diversité des structures, notamment au niveau des parties de discours et de la taille des textes, permet aussi de mieux appréhender la compositionnalité des textes. Dans (Mishra *et al.*, 2020), des méthodes de réduction de biais sont également proposées, comme la pseudonymisation des entités nommées. Le biais lié aux stéréotypes peut également être étudié en examinant la fréquence de N-grammes se rapportant à certains mots. Cela peut également permettre d'éviter une présence trop forte de biais textuels dans le corpus. Une manière de contrôler l'exactitude des textes est de limiter le bruit des données. Dans (Baldwin *et al.*, 2013), les auteurs comparent des textes issus de réseaux sociaux, pour évaluer le bruit des données. En analysant les mots hors vocabulaire et la grammaticalité des textes, ils évaluent d'une certaine manière l'exactitude des données collectées. De plus, ces auteurs proposent des méthodes pour nettoyer les textes afin de les rendre moins bruités.

5.2 Évaluation des images

La qualité d'un jeu de données de pré-entraînement pour la vision est d'abord déterminée par la qualité des images elles-mêmes, ce qui est compliqué de faire automatiquement. La taille des images

et les métadonnées associées peuvent fournir une aide pour cette évaluation.

Il est également important d'avoir une variabilité des images. On peut rechercher une couverture importante des possibles catégories, tels que des objets (Deng *et al.*, 2009), des scènes (Zhou *et al.*, 2017) ou des actions (Poppe, 2010). Pour cela, des ressources telles que Wikipédia ou WordNet (Miller, 1995) sont disponibles et peuvent être utilisées en lien avec des modèles détecteurs d'objets. En plus des objets représentés dans les images, il est intéressant d'obtenir diverses configurations d'images, comme des images iconiques et non iconiques, avec des dispositions et des nombres d'objets différents. Ceci peut être important pour favoriser la compréhension de la compositionnalité et une meilleure adaptation à divers domaines. Quant au bruit lié aux images, il peut être évalué en étudiant les catégories. En effet, une partie du biais d'un jeu de données peut provenir de l'utilisation de catégories bruitées, mal équilibrées ou contenant des stéréotypes (Prabhu & Birhane, 2020).

5.3 Évaluation de l'ancrage

La vérification de l'exactitude de l'ancrage requiert soit des annotations manuelles, soit l'utilisation d'un modèle vision-langage dont la performance n'est pas assurée. Ainsi, sur de grands jeux de données comme LAION, le modèle CLIP est utilisé pour mesurer la similarité entre texte et image. Cependant, le seuil de similarité requis semble fixé arbitrairement à partir d'observations d'annotateurs humains. Cela ne garantit pas la qualité de l'ancrage, notamment pour des différences texte-image fines qui peuvent ne pas être observées par CLIP. L'utilisation de modèles supplémentaires, comme les détecteurs d'objets, utilisés par les auteurs de Conceptual Captions, pourrait rendre l'évaluation plus robuste, et moins sensible aux spécificités d'un seul modèle. D'autre part, ces méthodes d'évaluation peuvent nuire à la diversité des données. En effet, CLIP est lui-même entraîné à partir de données collectées automatiquement, et peut reproduire les biais de ses données d'entraînement pendant l'évaluation. C'est pourquoi il serait intéressant d'utiliser conjointement plusieurs modèles entraînés sur des domaines ou tâches différents pour avoir différentes manières de juger l'ancrage texte-image des données.

Une autre problématique liée à l'évaluation de l'ancrage texte-image est le biais de l'annotation, qui se retrouve aussi en traitement automatique du langage (Geva *et al.*, 2019). En effet, dans de nombreux exemples de relation texte-image, celle-ci est subjective et dépend des annotateurs. Ils peuvent choisir de décrire différentes parties d'une image, en employant un langage varié, influencé par leurs diverses cultures. Il n'y a donc pas un unique ancrage textuel possible pour chaque image. Pour limiter ce biais, il est important d'obtenir des données provenant de sources variées, car un seul annotateur ne pourra être représentatif de cette subjectivité (Aroyo & Welty, 2015). Dans le cas d'annotations manuelles, il serait intéressant de donner différentes consignes d'annotation pour varier le type d'ancrage, avec diverses approches de description. Bien qu'il ne soit pas possible d'enlever tous les biais, la diversification des sources et la variabilité des textes et images peut aider à les atténuer.

Nous utilisons les méthodes développées dans cette section pour évaluer la qualité de deux jeux de données en Annexe A. Nous observons ainsi que LAION (Schuhmann *et al.*, 2021), qui est collecté automatiquement, montre une plus forte diversité de vocabulaire, mais une moindre compositionnalité que MS-COCO (Lin *et al.*, 2014), qui est annoté manuellement. Les annotations sont également moins descriptives et présentent plus d'information sur les métadonnées. Il serait

intéressant d'ajouter une étape d'analyse syntaxique lors du filtrage automatique des données afin de sélectionner des légendes plus descriptives et moins biaisées (moins de métadonnées et de noms propres). De plus, l'utilisation de détecteurs d'objets permettrait de sélectionner une plus grande diversité d'images en termes de nombre et catégories d'objets. Ainsi, une évaluation de la syntaxe et du vocabulaire des textes, ainsi que des objets contenus dans les images, permettrait d'améliorer la diversité et la compositionnalité des instances et de réduire l'impact du biais.

6 Discussions sur les problématiques éthiques

La collection d'un ensemble de données pour le pré-entraînement de modèles vision-langage peut soulever des questions éthiques, selon la manière dont les données sont obtenues et filtrées.

— L'utilisation de données provenant d'internet soulève la question du consentement lors de l'acquisition de ces données. Par exemple, Conceptual Captions (Sharma *et al.*, 2018) ou LAION (Schuhmann *et al.*, 2021) contiennent des images qui ne font pas l'objet d'une licence explicite. Certains collectent également des images et des textes contenant des informations personnelles sans demander le consentement des personnes concernées. Cela peut inclure des images provenant de sources problématiques et des données illégales.(Birhane *et al.*, 2021).

— Dans le cas d'ensembles de données annotées par des humains ou qui nécessitent des évaluations humaines, il est important de tenir compte de la rémunération des travailleurs et de leurs conditions de travail, par exemple de l'impact psychologique des contenus préjudiciables (Díaz *et al.*, 2022).

— Les données peuvent faire l'objet de biais nuisibles, en particulier dans le cas de données peu filtrées. L'une des formes de ces biais les plus courantes est l'absence de représentation de certains groupes sociaux (Zhao *et al.*, 2021), que l'on appelle 'biais de représentation'. Il peut avoir un impact important dans les applications en aval (Birhane *et al.*, 2021). Pour atténuer ce biais, il est important de tenir compte des sources de données visuelles et, lorsqu'il y a des annotateurs humains, de la diversité des expériences de ces annotateurs. (Díaz *et al.*, 2022).

7 Conclusion

Comme en traitement automatique des langues, l'utilisation de très grands jeux de données améliore très fortement le pré-entraînement des transformers vision-langage. Cependant, l'utilisation de plus grandes quantités de données conduit à un coût environnemental et économique non négligeable, ce qui rend leur démocratisation difficile. D'autres problèmes éthiques sont également soulevés, comme l'absence de consentement explicite lors de la collecte de ces données. Outre la taille des données, d'autres caractéristiques des jeux de pré-entraînement peuvent avoir un impact majeur sur la performance des modèles. En étudiant diverses analyses des transformers vision-langage, nous regroupons ces caractéristiques en variabilité, exactitude, compositionnalité et biais. Nous défendons un filtrage plus précis des grands jeux de données texte-image, pour mieux répondre à ces critères. Nous espérons qu'en accentuant l'importance de la qualité des données plutôt que leur quantité, le pré-entraînement des modèles vision-langage devienne plus efficace, et moins coûteux. Des études supplémentaires permettraient d'affiner les règles proposées.

8 Remerciements

Je voudrais remercier mes encadrants Benoit Favre et Stéphane Ayache, ainsi que les relecteurs pour leurs commentaires et suggestions.

Ces travaux ont bénéficié d'un accès aux ressources en IA de l'IDRIS au travers de l'allocation de ressources 2023-AD011013880 attribuée par GENCI. Ce travail a bénéficié d'une aide du gouvernement français au titre du Programme Investissements d'Avenir Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX (Institut Archimède AMX-19-IET-009)

Références

ABADJI J., SUAREZ P. O., ROMARY L. & SAGOT B. (2022). Towards a cleaner document-oriented multilingual crawled corpus. In *International Conference on Language Resources and Evaluation*.

AGRAWAL H., DESAI K., WANG Y., CHEN X., JAIN R., JOHNSON M., BATRA D., PARIKH D., LEE S. & ANDERSON P. (2019). Nocaps : Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, p. 8948–8957.

ANTOL S., AGRAWAL A., LU J., MITCHELL M., BATRA D., ZITNICK C. L. & PARIKH D. (2015). Vqa : Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, p. 2425–2433.

AROYO L. & WELTY C. (2015). Truth is a lie : Crowd truth and the seven myths of human annotation. *AI Magazine*, **36**(1), 15–24.

BALDWIN T., COOK P., LUI M., MACKINLAY A. & WANG L. (2013). How noisy social media text, how diffrnt social media sources ? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 356–364.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & 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*, p. 610–623.

BIRHANE A., PRABHU V. U. & KAHEMBWE E. (2021). Multimodal datasets : misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv :2110.01963*.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

BYEON M., PARK B., KIM H., LEE S., BAEK W. & KIM S. (2022). Coyo-700m : Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

CHANGPINYO S., SHARMA P., DING N. & SORICUT R. (2021). Conceptual 12M : Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

CHEN Y.-C., LI L., YU L., EL KHOLY A., AHMED F., GAN Z., CHENG Y. & LIU J. (2019). Uniter : Learning universal image-text representations.

DENG J., DONG W., SOCHER R., LI L.-J., LI K. & FEI-FEI L. (2009). Imagenet : A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, p. 248–255 : Ieee.

DEVLIN J., CHANG M.-W., LEE K. & 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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DÍAZ M., KIVLICHAN I. D., ROSEN R., BAKER D., AMIRONESEI R., PRABHAKARAN V. & DENTON E. L. (2022). Crowdworksheets : Accounting for individual and collective identities underlying crowdsourced dataset annotation. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J. & HOULSBY N. (2021). An image is worth 16x16 words : Transformers for image recognition at scale. In *International Conference on Learning Representations*.

GEVA M., GOLDBERG Y. & BERANT J. (2019). Are we modeling the task or the annotator ? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv :1908.07898*.

GOYAL Y., KHOT T., SUMMERS-STAY D., BATRA D. & PARIKH D. (2017). Making the v in VQA Matter : Elevating the Role of Image Understanding in Visual Question Answering. p. 6904–6913.

HENDRICKS L. A., BURNS K., SAENKO K., DARRELL T. & ROHRBACH A. (2018). Women also Snowboard : Overcoming Bias in Captioning Models. p. 771–787.

HENDRICKS L. A., MELLOR J. F. J., SCHNEIDER R., ALAYRAC J.-B. & NEMATZADEH A. (2021). Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, **9**, 570–585.

HENDRICKS L. A. & NEMATZADEH A. (2021). Probing Image-Language Transformers for Verb Understanding. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3635–3644, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.318.

HENDRYCKS D., LIU X., WALLACE E., DZIEDZIC A., KRISHNAN R. & SONG D. X. (2020). Pretrained transformers improve out-of-distribution robustness. In *Annual Meeting of the Association for Computational Linguistics*.

HONNIBAL M. & MONTANI I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

HOVY D. & PRABHUMOYE S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, **15**(8), e12432.

HUDSON D. A. & MANNING C. D. (2019). Gqa : a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv :1902.09506*, **3**(8), 1.

KAY M., MATUSZEK C. & MUNSON S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, p. 3819–3828.

KREUTZER J., CASWELL I., WANG L., WAHAB A., VAN ESCH D., ULZII-ORSHIKH N., TAPO A., SUBRAMANI N., SOKOLOV A., SIKASOTE C. *et al.* (2022). Quality at a glance : An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, **10**, 50–72.

KRISHNA R., ZHU Y., GROTH O., JOHNSON J., HATA K., KRAVITZ J., CHEN S., KALANTIDIS Y., LI L.-J., SHAMMA D. A., BERNSTEIN M. & FEI-FEI L. (2016). Visual genome : Connecting language and vision using crowdsourced dense image annotations.

LI J., LI D., XIONG C. & HOI S. C. H. (2022). Blip : Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

LI J., SELVARAJU R., GOTMARE A., JOTY S., XIONG C. & HOI S. C. H. (2021). Align before fuse : Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, **34**, 9694–9705.

LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P. & ZITNICK C. L. (2014). Microsoft coco : Common objects in context. In *Computer Vision–ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, p. 740–755 : Springer.

LIU F., BUGLIARELLO E., PONTI E. M., REDDY S., COLLIER N. & ELLIOTT D. (2021). Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 10467–10485, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.818.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

MISHRA S., ARUNKUMAR A., SACHDEVA B., BRYAN C. & BARAL C. (2020). Dqi : Measuring data quality in nlp. *arXiv preprint arXiv :2005.00816*.

NIKOLAUS M., SALIN E., AYACHE S., FOURTASSI A. & FAVRE B. (2022). Do vision-and-language transformers learn grounded predicate-noun dependencies ? *arXiv preprint arXiv :2210.12079*.

ORDONEZ V., KULKARNI G. & BERG T. (2011). Im2text : Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, **24**.

OTTO C., SPRINGSTEIN M., ANAND A. & EWERTH R. (2019). Understanding, categorizing and predicting semantic image-text relations. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, p. 168–176.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A. *et al.* (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv :2203.02155*.

POPPE R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, **28**(6), 976–990.

PRABHU V. U. & BIRHANE A. (2020). Large image datasets : A pyrrhic win for computer vision ? *arXiv preprint arXiv :2006.16923*.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J. *et al.* (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, p. 8748–8763 : PMLR.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I. *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9.

REN S., HE K., GIRSHICK R. & SUN J. (2015). Faster r-cnn : Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, **28**.

SALIN E. (2022). Etude de la compréhension spatiale multimodale des modèles transformers vision-langage. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, p. 181–187 : CNRS.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

SCHUHMANN C., BEAUMONT R., VENCU R., GORDON C., WIGHTMAN R., CHERTI M., COOMBES T., KATTA A., MULLIS C., WORTSMAN M. *et al.* (2022). Laion-5b : An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv :2210.08402*.

SCHUHMANN C., VENCU R., BEAUMONT R., KACZMARCZYK R., MULLIS C., KATTA A., COOMBES T., JITSEV J. & KOMATSUZAKI A. (2021). Laion-400m : Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv :2111.02114*.

SCHWARTZ R., DODGE J., SMITH N. A. & ETZIONI O. (2020). Green ai. *Communications of the ACM*, **63**(12), 54–63.

SHARMA P., DING N., GOODMAN S. & SORICUT R. (2018). Conceptual captions : A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

SINGH A., GOSWAMI V. & PARIKH D. (2020). Are we pretraining it right ? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv :2004.08744*.

STRUBELL E., GANESH A. & MCCALLUM A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv :1906.02243*.

SUAREZ P. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures.

SUHR A., ZHOU S., ZHANG A., ZHANG I., BAI H. & ARTZI Y. (2019). A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 6418–6428, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1644.

SUN C., SHRIVASTAVA A., SINGH S. & GUPTA A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, p. 843–852.

TAN H. & BANSAL M. (2019). LXMERT : Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 5100–5111, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1514.

TOUVRON H., CORD M., DOUZE M., MASSA F., SABLAYROLLES A. & JÉGOU H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, p. 10347–10357 : PMLR.

WENZEK G., LACHAUX M.-A., CONNEAU A., CHAUDHARY V., GUZMÁN F., JOULIN A. & GRAVE E. (2019). Ccnet : Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv :1911.00359*.

YANG J., DUAN J., TRAN S., XU Y., CHANDA S., CHEN L., ZENG B., CHILIMBI T. & HUANG J. (2022). Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 15671–15680.

ZHAI X., KOLESNIKOV A., HOULSBY N. & BEYER L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 12104–12113.

ZHAO D., WANG A. & RUSSAKOVSKY O. (2021). Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 14830–14840.

ZHAO J., WANG T., YATSKAR M., ORDONEZ V. & CHANG K.-W. (2017). Men also like shopping : Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv :1707.09457*.

ZHOU B., LAPEDRIZA A., KHOSLA A., OLIVA A. & TORRALBA A. (2017). Places : A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, **40**(6), 1452–1464.

ZHU Y., GROTH O., BERNSTEIN M. & FEI-FEI L. (2016). Visual7w : Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4995–5004.

A Comparaison de MS COCO et LAION-400

En utilisant les méthodes d'évaluation développées dans la section 5, nous comparons deux corpus collectés manuellement et automatiquement : MS COCO (Lin *et al.*, 2014) et LAION-400 (Schuhmann *et al.*, 2021). Nous étudions un sous-ensemble S de 1000 instances de chaque jeu de données et comparons les jeux de données sur plusieurs mesures :

— Variabilité du vocabulaire : En appelant d la taille du dictionnaire de S et n le nombre de mots de S, nous calculons $V = d/n$.
On obtient $V_{COCO} = 0.04$ et $V_{Laion} = 0.10$.
LAION montre ainsi une plus grande diversité de vocabulaire que MS COCO.

— Biais du vocabulaire : Afin d'obtenir plus d'information sur le vocabulaire utilisé par ces jeux de données, nous étudions quels groupes de mots sont les plus utilisés. Cela peut permettre d'étudier le biais de chaque sous-ensemble S. En particulier, après avoir éliminé les mots vide, nous regardons quel groupe G de deux mots se retrouve le plus fréquemment dans une même instance S.
On obtient $G_{COCO} =$ (group, people) et $G_{Laion} =$ (stock, photo).
Cela semble montrer que COCO contient des textes descriptifs, alors que les textes de Laion contiennent également beaucoup de métadonnées.

— Nombre d'objets : Le nombre d'objets dans une image permet d'avoir plus d'information sur la complexité de la composition de cette image. Nous calculons à laide d'un détecteur d'objet Faster RCNN (Ren *et al.*, 2015). le nombre d'objets par image de S, et on rapporte la médiane M et le troisième quartile Q.
On obtient $M_{COCO} = 4$, $Q_{COCO} = 8$ et $M_{Laion} = 1$, $Q_{Laion} = 2$.
On observe que les images de COCO contiennent plus d'objets que celles de LAION. Cette méthode a cependant des limites, notamment parce que les images du jeu de données LAION sont peu semblables à celles utilisées pour l'entraînement du détecteur d'objet.

— Syntaxe : Étudier la syntaxe d'un texte peut également nous donner plus d'information sur la qualité de celui-ci. Pour ce faire, nous utilisons les outils de Spacy (Honnibal & Montani, 2017). Nous observons les deux plus fréquentes étiquettes de partie de discours $P1$ et $P2$. On obtient $P1_{COCO}$ = Nom et $P2_{COCO}$ = Déterminant, $P1_{Laion}$ = Nom et $P2_{Laion}$ = Nom propre.
La prévalence de noms propres dans le jeu de données LAION peut augmenter le risque de biais.