Faire écrire son papier par un programme d IA rêve ou realite

Philippe Very, Tournois Alice

Faire écrire son papier par un programme d'intelligence artificielle : rêve ou réalité ?

L'intelligence artificielle (IA) pénètre la plupart des disciplines scientifiques, parfois depuis plusieurs années comme en médecine ; elle émerge depuis peu dans le secteur de la gestion. Plusieurs chercheurs ont montré comment l'IA pourrait modifier la conduite d'un projet de recherche et sa publication (Benavent, 2016 ; Baumard, 2019 ; Very et Cailluet, 2019).

Le sujet abordé lors de la communication consistera à présenter deux programmes récents d'intelligence artificielle (IA) en *open source*. Le premier programme s'appelle Elicit. Il est destiné à aider le chercheur à mener une analyse de littérature. Le second, ChatGPT, vise à écrire un résumé des connaissances acquises sur n'importe quel sujet, de recherche ou autre.

Lors de la présentation, nous comptons analyser les points forts et les risques associés à ces programmes à partir de démonstrations « *live* » et à partir d'expérimentations que nous avons menées. Les résultats obtenus, conjugués avec les résultats d'expériences effectuées par d'autres chercheurs, renseignent sur la façon dont ces programmes peuvent contribuer à l'écriture scientifique. Nous joignons dans l'annexe les connaissances accumulées par d'autres chercheurs sur Chat GPT.

Références principales

Baumard, P. (2019) Quand l'intelligence artificielle théorisera les organisations, Revue Française de Gestion, No 285, p 135-159.

Benavent, C. (2016). « Big Data, algorithmes et marketing : rendre des comptes », Statistique et Société, Vol. 4, N° 3, 25-35.

Very, P. & Cailluet, L. (2019) Intelligence Artificielle et Recherche en Gestion, Revue Française de Gestion, No 285, p 120-134.

Annexe : Présentation et discussion de Chat-GPT – revue de littérature

1) Qu'est-ce que c'est ?

Chat-GPT est une application de « traitement automatique des langues » (en Anglais NLP : Natural Language Processing) qui est capable de générer un texte en réponse à une demande de son utilisateur.

Ces programmes ingurgitent de grandes masses de données sur le langage et apprennent à formuler une phrase, puis un paragraphe, puis un texte cohérent à partir de ces données. Ils ne sont pas nouveaux : de tels programmes sont utilisés depuis 40 ans pour la correction orthographique ou la traduction automatique.

La technique de base est plutôt simple : il s'agit il s'agit de prédire le mot qui est le plus probable en fonction des deux ou trois mots précédents. Le programme analyse dont de grandes quantités de texte et de séquences dans ces textes, et encode les séquences.

Les nouveaux modèles de langage comme Chat-GPT ou BERT se distinguent de leurs prédécesseurs par l'ampleur des données ingurgitées et la sophistication de la technique de base. D'après ses concepteurs, Chat-GPT a appris à partir de 175 milliards de paramètres (contextes d'emploi des mots) et plusieurs centaines de milliards de mots. Poibeau (2021) souligne que c'est bien plus de données qu'un être humain pourrait percevoir durant toute sa vie.

La technique utilisée par Chat-GPT et les programmes analogues est le modèle de « transformer » qui ne se contente pas d'analyser les 2 mots précédents. Le modèle analyse des éléments linguistiques dans le contexte de la phrase, et encode des informations sur les mots et leur contexte d'utilisation. Ces techniques se classent dans la catégorie des « apprentissages non supervisés » (*unsupervised learning)*. Un apprentissage non supervisé signifie que le programme fait ses propres inférences sans intervention humaine. Le modèle utilise le mimétisme pour prédire : par exemple, tel mot est le plus probable dans cette phrase.

Ces applications génèrent des textes présentant une syntaxe d'excellente qualité et en général cohérents. Les textes peuvent être rédigés de façon sérieuse, drôle ou poétique selon la demande. Ces modèles sont aussi capables de répondre à des questions ou de faire des opérations mathématiques simples.

2) Les dangers liés à ces programmes

Bender et al. (2021) font une synthèse des 4 principaux dangers liés à ces programmes.

1. Coûts environnementaux et financiers
   Ces programmes requièrent une grande puissance de traitement informatique, et consomment donc beaucoup d'électricité. Plus il y a de données ingurgitées, plus le modèle consomme (Strubell et al. 2019). L'apprentissage des modèles les plus puissants comme Chat-GPT émettrait a minima 60 fois plus de CO2 qu'un humain durant toute sa vie. Bender et al. (2021) concluent que seules es organisations riches peuvent se permettre d'entraîner de tels modèles alors que ce sont surtout les populations pauvres qui souffrent le plus de la

dégradation de la planète. Very et Cailluet (2019) soulignaient aussi le fait que le développement de tels modèles d'intelligence artificielle pourrait accroître la fracture entre les riches capables de se les payer et les pauvres n'y ayant pas accès.

2. Biais non contrôlables

   Chat-GPT et les programmes analogues utilisent essentiellement des données récoltées via Internet pour leur apprentissage. Or, il y a de fortes chances que ces données contiennent des biais. Par exemple, GPT-2 a été nourri, entre autres sources, par des données émanant de Reddit, Wikipedia et Twitter. Une étude menée par Pew Internet Research en 2016 a montré que 67% des utilisateurs US de Reddit étaient des hommes, et 64% avaient entre 18 et 29 ans (Pew Internet Research, 2016). Barera (2020) a montré que seulement 8 à 15% des auteurs de Wikipedia étaient de sexe féminin. De tels scores peuvent créer des biais dans l'apprentissage, biais similaires à ceux involontairement engendrés par des auteurs humains de programmes d'IA (O'Neil, 2016). En faisant avaler toutes sortes de données sans contrôle à ces programmes, il y a des chances que les textes générés puissent contenir des biais de sexe, de race ou autre. Ces programmes risquent aussi de ne pas identifier les subtilités de langage, les spécificités culturelles des peuples et pays, et de tendre à homogénéiser le langage.

   Et même lorsqu'il y a filtrage des données, la pratique actuelle qui consiste à éliminer les documents contenant des mots défendus privilégie au final la vision dominante. En sélectionnant de grandes quantités de données Internet jugées acceptables, il y a donc risque de perpétuer les points de vue dominants, d'accroître les déséquilibres de pouvoir et de renforcer les inégalités.

   Il y a donc des biais possibles liés aux sources de données et aux participants sur internet, ainsi que des biais possibles liés aux éventuels filtres mis en place.

3. Opportunisme des investisseurs

   Les programmes tels Chat-GPT « ne savent pas qu'ils ne savent pas » (Suchanek et Varoquaux, 2022). Ils manipulent le langage. Or ces programmes sont financés par Microsoft (Chat-GPT) ou Google (BERT). Pour Bender et al. (2021), ces géants de la Tech sont des entreprises à but lucratif et voudront prochainement rentabiliser leur investissement. Leur effort actuel de recherche est influencé par leur opportunisme : ils font avaler de plus en plus de données pour avoir des résultats plus précis, mais ceci est fait au détriment de la pertinence des résultats qui pourrait être obtenue avec des bases de données plus petites mais mieux sélectionnées.

4. Illusion de vérité

   Le quatrième problème posé par ces programmes est lié à leur forte capacité d'imitation de l'écriture humaine. Il y a donc un fort risque de les utiliser pour tromper la foule. En témoigne le cas de l'étudiant américain qui a produit des conseils d'auto-assistance et de productivité générés par l'IA sur un blog qui est devenu viral (MIT Technology Review, 2020). Lors de cette expérience, des suiveurs du blog ont farouchement défendu le blog face à des internautes suspectant des recommandations générées par l'IA. Il y a donc un vrai risque de manipulation des foules.

3) Conclusion : comment utiliser en recherche

Les programmes comme Chat GPT génèrent dont une dette de documentation dans le sens où les données sont à la fois non documentées et trop volumineuses pour être documentées *a posteriori*. Sans documentation, on ne peut pas essayer de comprendre les données d'entraînement afin d'atténuer certains problèmes attestés ou de générer des problèmes inconnus. La solution préconisée par Bender et al. (2021) consiste à investir des ressources importantes dans la conservation et la documentation des données d'entraînement. Jo et al. (2020) appellent à utiliser les méthodes d'archivage pour la collecte des données historique (*archival history data collection methods*). Birhane et Prabhu (2021) et Benjamin [2019] appellent à une méthodologie de collecte de données intégrant la notion de justice : « *nourrir des programmes AI avec la beauté, la laideur et la cruauté du monde, mais en s'attendant à ce qu'il révèle seulement la beauté, est un fantasme*. » (Benjamin, 2019 : p.1541).

Idéalement, les concepteurs de tels programmes d'IA devront décrire les usages pour lesquels leurs modèles ont été élaborés Il leur faudra fournir une documentation complète sur les données utilisées dans la construction du modèle, y compris leurs motivations sous-jacentes au processus de sélection et de collecte des données. Cette documentation devrait indiquer les objectifs, les valeurs et les motivations des chercheurs ; elle devrait aussi indiquer les utilisateurs et parties prenantes qui risquent d'être négativement affectés par des erreurs de modèle ou une mauvaise utilisation. (Bender et al., 2021).

Ces recommandations concernent essentiellement les créateurs de ces programmes. Concernant les chercheurs utilisateurs, la plus grande prudence est donc de mise, compte tenu des dangers évoqués. Il faudra certainement attendre la seconde génération de programmes NLP, sélectionner celui qui sera construit sur la base de données la plus pertinente pour la recherche et l'utiliser en connaissance des biais possibles.

Références

Barera, M. (2020). *Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia*. Accessible à http://hdl.handle.net/10106/29572

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*. pp. 610-623.

Benjamin., R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.

Birhane, A., & Prabhu, V.U. (2021). Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp 1537–1547.

Jo, E. S., & Gebru, T. (2020.) Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. pp 306–316.

MIT Technology Review, (2020*). A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it*. Accessible à https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/

O'Neil, C. (2016) *Weapons of Math Destruction*. Crown Books (USA).

Pew Internet Research (2016). https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male, young-and-digital-in-their-news-preferences/

Poibeau, T. (2021). Quand l'IA prend la parole : des prouesses aux dangers. *The Conversation*, 25 Janvier 2021. Accessible à https://theconversation.com/quand-lia-prend-la-parole-des-prouesses-aux-dangers-153495

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

Suchanek, F., & Varoquaux, G. (2022) Beau parleur comme une IA. *The Conversation*. 26 décembre 2022. Accessible à https://theconversation.com/beau-parleur-comme-une-ia-196084

Very, P., & Cailluet L. (2019) Intelligence Artificielle et Recherche en Gestion, *Revue Française de Gestion*. No 285. pp 120-134.