Un rapport de

cdt | Research

RÉSUMÉ EXÉCUTIF - FRANÇAIS

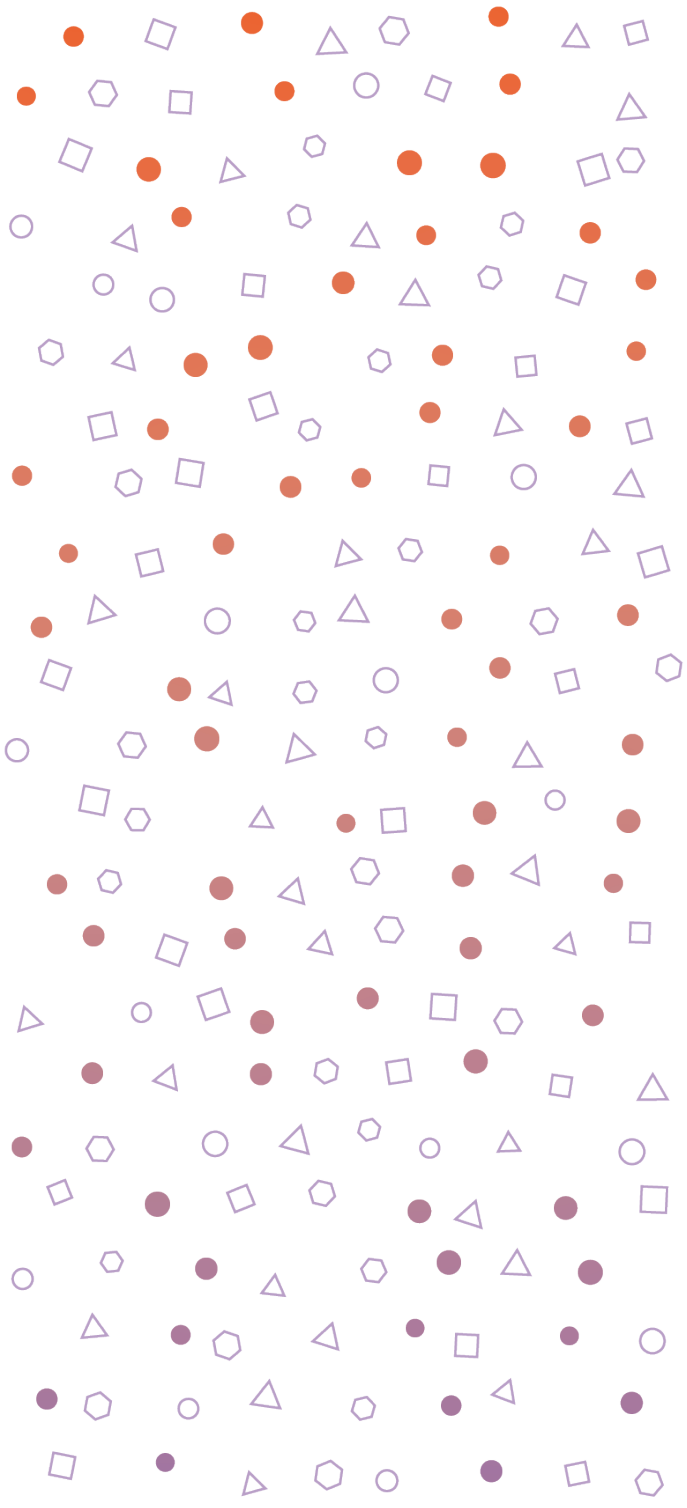
Lost in Translation

les modèles de langue de grande taille dans l'analyse de contenu non-anglophone

Gabriel Nicholas
Aliya Bhatia

mai 2023

Résumé Exécutif



Pour la plupart des gens dans le monde, internet est la source principale d'informations, d'opportunités économiques et de communauté. Cependant, l'usage croissant des systèmes automatisés dans nos interactions en ligne - tels que les chatbots, les modérateurs automatiques de contenus et les moteurs de recherche - sont conçus à l'origine en anglais et fonctionnent bien plus efficacement dans cette langue que dans l'une des 7000 autres langues du monde.

Ces dernières années, les modèles de langue de grande taille sont devenus le vecteur dominant pour élaborer des systèmes d'intelligence artificielle afin d'analyser et de générer du texte en ligne. Ces modèles sont néanmoins principalement conçus en anglais et pour la langue anglaise. Un modèle linguistique de grande taille (par exemple, GPT-4 de Open AI, LLaMa de Meta et PaLM de Google) est un algorithme d'apprentissage automatique capable d'analyser de très vastes quantités de données textuelles afin d'apprendre quels mots et phrases apparaissent fréquemment à proximité les uns des autres et dans quels contextes. Les modèles linguistiques de grande taille peuvent exécuter une grande variété de tâches dans différents domaines. Ils sont surtout connus du grand public pour concevoir des chatbots, comme ChatGPT, mais les chercheurs et les entreprises technologiques les utilisent aussi pour effectuer des tâches d'analyse de contenu, telles que l'analyse des émotions, les résumés de textes et la détection de discours de haine. Google, Meta, Microsoft et d'autres entreprises ont d'ores et déjà incorporé les grands modèles linguistiques dans les fonctions primaires de leurs produits, comme la modération de contenu et les moteurs de recherche. D'autres fournisseurs de services pourraient bientôt les incorporer dans des systèmes de prise de décision automatisée, tels que la lecture automatique des CVs.

Depuis peu, néanmoins, les chercheurs et les entreprises technologiques tentent d'élargir les capacités des grands modèles de langage à des langues autres que l'anglais en construisant des modèles linguistiques multilingues. A défaut d'être entraînés avec des données textuelles d'une langue unique, les modèles multilingues le sont sur base de textes de dizaines ou centaines de langues différentes à la fois. Les chercheurs postulent que les modèles linguistiques multilingues déduisent les connexions entre langues, ce qui leur permet d'appliquer des associations de mots et des règles grammaticales sous-jacentes apprises

avec des langues disposant de davantage de données textuelles (en particulier l'anglais) pour l'entraînement à celles qui en possèdent moins. Pour certaines applications, les modèles linguistiques multilingues opèrent d'ailleurs mieux que ceux entraînés sur une seule langue. Par exemple, les performances d'un modèle entraîné sur une grande quantité de textes en de nombreuses langues, dont l'hindi, contextualiseraient mieux l'hindi qu'un modèle linguistique entraîné uniquement sur des textes en hindi.

Les modèles linguistiques multilingues fournissent aux entreprises technologiques le moyen d'utiliser leurs systèmes d'intelligence artificielle dans de nombreuses langues en une seule fois, et certaines ont déjà commencé à les intégrer dans leurs produits. Les fournisseurs de services en ligne, en particulier, ont déployé des modèles linguistiques multilingues pour modérer leur contenu: Meta utilise un modèle linguistique multilingue pour détecter les contenus néfastes sur ses plateformes dans 100 langues ; l'API Perspective de Alphabet pour trouver les contenus toxiques dans 18 langues différentes ; le site de rencontres Bumble pour détecter les messages non sollicités à caractère sexuel dans le monde entier.

Les technologues ont recours aux modèles linguistiques multilingues pour élaborer des modèles dans des langues qui ne disposent pas de suffisamment de données textuelles. Pour entraîner leur modèles linguistiques, les technologues sont confrontés à une grande disparité entre les différentes langues en termes de disponibilité des ressources textuelles, c'est-à-dire en volume, qualité et variété de données textuelles. L'anglais est de très loin la langue disposant de la plus grande quantité de ressources textuelles, suivie de près par l'espagnol, le chinois, l'allemand, et une poignée d'autres langues, permettant la construction aisée de modèles linguistiques pour ces langues. Pour les langues à ressources intermédiaires, avec des ensembles de données moins nombreux mais de grande qualité telles que le russe, l'hébreu et le vietnamien, et les langues à faibles ressources, ne disposant quasiment d'aucun jeux de données disponibles pour l'entraînement, comme l'amharic, le cherokee, et le créole haïtien, disposent de trop peu de données textuelles s pour former leurs propres modèles de langue de grande taille. Les données des langues à faibles ressources textuelles sont par ailleurs souvent de mauvaise qualité. Soit elles sont soit mal traduites, voire incompréhensibles, tirées d'internet, soit elles sont limitées à des sources aux domaines étroits telles que les textes religieux et Wikipedia. Cet écart dans la disponibilité des données entre les langues est connu sous le nom de déficit de ressources.

Les modèles linguistiques multilingues sont conçus pour combler ces lacunes dans la disponibilité des données disponibles, en déduisant les connexions sémantiques et grammaticales entre les langues de ressources textuelles élevées et déficitaires, ce qui permet aux premières d'amorcer les secondes. Toutefois, cette architecture soulève ses propres problèmes. Les modèles linguistiques multilingues sont encore généralement conçus de façon disproportionnée sur des textes en anglais et finissent donc par transférer des valeurs et des hypothèses encodées en anglais dans d'autres contextes linguistiques où elles n'ont pas forcément leur place. Par exemple, un modèle multilingue peut associer le mot "colombe" dans toutes les langues au mot "paix", même si le mot pour colombe en basque ("uso") peut être une insulte. La disparité des données disponibles signifie également que les modèles linguistiques multilingues fonctionnent beaucoup mieux dans les langues à ressources textuelles élevées et les langues apparentées que dans les langues à ressources faibles. Les développeurs de modèles tentent parfois de combler ces déficits par

des textes traduits automatiquement, mais les erreurs de traduction peuvent aggraver davantage la représentation erronée de la langue. Par ailleurs, lorsque les modèles linguistiques multilingues échouent, les liens non intuitifs qu'ils établissent entre les langues peuvent rendre ces problèmes plus difficiles à identifier, à diagnostiquer et à résoudre.

L'utilisation courante des grands modèles linguistiques dans l'analyse de contenu soulève d'autres questions. Les linguistes informaticiens soutiennent que les grands modèles linguistiques disposent d'une capacité limitée pour analyser des formes d'expression absentes des données utilisées pour leur apprentissage, ce qui signifie qu'ils peuvent avoir des difficultés à fonctionner correctement dans des contextes nouveaux. Ils peuvent par ailleurs reproduire tous les biais présents dans les données utilisées pour leur apprentissage. Souvent, les données textuelles sont prélevées sur internet, ce qui signifie que les modèles linguistiques de grande taille peuvent encoder et renforcer des opinions dominantes exprimées en ligne.

Les entreprises, les chercheurs et les gouvernements ont chacun un rôle à jouer pour protéger la population des possibles dangers des systèmes d'analyse de contenu alimentés par les modèles linguistiques multilingues. Pour mieux informer le public, les entreprises qui déploient des grands modèles linguistiques devraient toujours être transparentes sur la manière dont elles les utilisent et dans quelles langues. Elles devraient également octroyer des responsabilités limitées aux grands modèles linguistiques qu'elles déploient, ainsi que des canaux adéquats pour permettre une supervision humaine.

En parallèle, les chercheurs et les bailleurs de fonds de la recherche devraient investir dans des actions visant à améliorer l'utilisation et les performances des modèles linguistiques en des langues autres que l'anglais, en particulier afin de réduire les dysfonctionnements qui ont des répercussions négatives sur les locuteurs de langues à faibles ressources textuelles. La meilleure façon d'y parvenir est de soutenir les communautés de chercheurs spécialisés en linguistique, qui peuvent promouvoir le cycle vertueux de la collecte de données, de la conservation des jeux de données, de l'entraînement des modèles de langage, de la publication et de la création d'applications. Les locuteurs des langues locales et les experts du contexte doivent participer à chaque étape de ce processus, conserver les données et évaluer les modèles linguistiques déployés par les grands fournisseurs de services en ligne opérant dans le monde entier.

Enfin, les gouvernements doivent être prudents quant à la manière dont ils utilisent ou encouragent l'utilisation des modèles linguistiques de grande taille. Les grands modèles ne devraient jamais alimenter des systèmes utilisés pour prendre des décisions délicates sans supervision, dans des domaines tels que le statut migratoire ou la santé. Les gouvernements ne devraient pas non plus mandater ou requérir par inadvertance via la législation l'utilisation de systèmes pilotés par des grands modèles linguistiques pour modérer les contenus de services en ligne. Au contraire, il leur incomberait de réunir les différentes parties prenantes pour qu'elles s'accordent sur les normes et garde-fous qui devraient entourer le développement et le déploiement de modèles linguistiques de grande taille.

Les modèles linguistiques de grande taille en général, et les modèles linguistiques multilingues en

particulier, ont le potentiel de créer de nouvelles opportunités économiques et d'améliorer le web pour tout un chacun. Cependant, l'utilisation erronée ou excessive de ces technologies représente une réelle menace pour les droits humains, tels que le droit à la liberté d'expression, en supprimant par exemple, sans raison valable, un message publié sur les réseaux sociaux, ou encore le droit à ne pas subir de discriminations si la profession d'une personne ou une demande de visa sont mal interprétées. Les modèles linguistiques multilingues, en particulier, peuvent involontairement renforcer l'anglo-centrisme qu'ils sont supposés combattre. Compte tenu de ces limites, les entreprises technologiques, les chercheurs et les gouvernements doivent tenir compte des éventuels risques pour les droits humains et civils quand ils étudient, fournissent, développent ou utilisent des grands modèles linguistiques pour alimenter des systèmes, en particulier lorsque ces derniers sont utilisés pour mettre à disposition des informations sensibles ou qu'ils jouent un rôle dans des décisions qui affectent l'accès des personnes à des opportunités économiques, à la liberté, ou à d'autres intérêts et droits importants.

A report from

cdt | Research

Read the full report at *cdt.org*.

CONTRIBUTIONS BY

Samir Jain, Mallory Knodel, Emma Llansó, Michal Luria, Nathalie Maréchal, Dhanaraj Thakur, and Caitlin Vogus.

ACKNOWLEDGEMENTS

We thank Pratik Joshi, Sebastin Santy, and Aniket Kesari for their invaluable feedback on the technical aspects of this report. We also thank Jacqueline Rowe, Damini Satija, and Ángel Díaz for their insightful comments and suggestions. All views in this report are those of CDT.

The translation of our executive summary is made possible by Global Voices Translations and with the help of Iverna McGowan, Maria Villamar, Ophélie Stockhem, and Tomás Pomar.

This work is made possible through a grant from the John S. and James L. Knight Foundation.

Suggested Citation: Nicholas, G. and Bhatia, A. (2023) Lost in Translation: Large Language Models in Non-English Content Analysis. Center for Democracy & Technology. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1996, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C., and has a Europe Office in Brussels, Belgium.

GABRIEL NICHOLAS

Research Fellow at the Center for Democracy & Technology.

ALIYA BHATIA

Policy Analyst, Free Expression Project at the Center for Democracy & Technology.

CENTER FOR
DEMOCRACY
& TECHNOLOGY