Petits oublis, grands effets : le silençage des communauté linguistiques minorisées dans le TAL et ses conséquences

Mélanie Jouitteau, Loïc Grobol

Petits oublis, grands effets : le silençage des communauté linguistiques minorisées dans le TAL et ses conséquences

Mélanie Jouitteau
IKER, UMR 5478, CNRS,
Université de Pau et des Pays de l'Adour,
et Université Bordeaux-Montaigne
melanie.jouitteau@iker.cnrs.fr

Loïc Grobol
MoDyCo, CNRS,
et Université Paris Nanterre ;
Lattice, ENS, CNRS,
et Université Sorbonne Nouvelle
lgrobol@parisnanterre.fr

Un ensemble de langues à corpus restreint fait les frais d'une stratégie publicitaire de façade (« *diversity washing* »). Il s'agit typiquement des langues de niveau de développement numérique intermédiaire. Ce ne sont ni celles qui n'ont qu'une poignée de corpus, ni la cinquantaine de langues les plus développées numériquement, protégées par des cultures hégémoniques et des appareils d'État.

Pour ces langues de niveau de développement numérique intermédiaire, il existe sur le Web quelques données, qui peuvent être moissonnées, pour peu que l'on dispose d'un reconnaisseur de langue, même rudimentaire. Vu de loin, si on ne les examine pas sérieusement, ces données ont la masse critique nécessaire au développement d'outils de TAL. De grandes entreprises internationales annoncent ainsi développer pour ces langues des outils fonctionnels, ce qui leur permet de rassurer les sociétés sur leur implication et leur conscience sociale. Ces entreprises ne mettent cependant en place aucune démarche qualité en direction des locuteurs et utilisateurs et, de façon assez peu surprenante, en l'absence de remontées d'information venant des locuteurs eux-mêmes et d'évaluations construites avec leur expertise, leurs outils dysfonctionnent sans signal d'alarme.

1 La traduction breton → français

Nous proposons comme exemple concret de ce défaut de prise en charge le cas de la traduction automatique breton-français. Le tableau 1 rapporte les performances du système de traduction automatique multilingue m2m100 (Fan et al., 2021), annoncé comme traitant 100 langues, dont le breton et le français, entraîné à l'aide de corpus parallèles collectés automatiquement et ne proposant pas d'évaluation dans toutes les langues concernées (et notamment pas d'évaluation pour le breton).

Nous évaluons ce système sur un corpus de quelques centaines de phrases inédites conçu pour l'occasion. Nous améliorons ce modèle en poursuivant son entraînement sur deux corpus de qualité, mis au point par ou avec la collaboration étroite de brittophones.

Tab. 1 : Performances d'un système de traduction automatique breton→français entraîné sur des données multilingues pour lesquelles le breton est mal aligné (m2m100-418M), puis en ajoutant un corpus général de breton de qualité et de taille moyenne (OPAB, Tyers (2009)), et enfin en ajoutant un corpus petit mais de très haute qualité et diversité (ARBRES, Jouitteau (2009-2024)). Les scores sont calculés avec les paramètres par défaut de SacreBLEU (Post, 2018)

Modèle	BLEU	ChrF++	TER
m2m100-418M	0.58	11.85	114.49
+OPAB	30.01	50.16	55.37
+ARBRES	37.68	56.99	48.65

Le premier constat est la qualité catastrophique du système, pour une langue qu'il est pourtant annoncé supporter. L'entraînement sur les données de Tyers (2009) améliore drastiquement les performances, et l'ajout de celles de Jouitteau (2009-2024) encore davantage. Ces données étaient pourtant disponibles publiquement au moment de l'élaboration du système initial et, au moins pour Tyers (2009), bien connues, ayant déjà été utilisées par plusieurs travaux de recherche. Il aurait ainsi suffi d'entrer en contact avec la communauté linguistique à ce moment pour en disposer, et s'assurer de ne pas diffuser un système aux performances aussi catastrophiques.

Du point de vue qualitatif, pour ne donner qu'un seul exemple, la phrase « *Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ* », dont une traduction possible en français est « *La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.* » est traduite ainsi par ces différents modèles :

m2m100 « *C'est le cas d'un homme qui a laissé le coucher, et qui a laissé le coucher.* »

+OPAB « *La langue dans laquelle elle fait un homme est un monde dans lequel elle vit et s'efforce.* »

+OPAB+ARBRES « *La langue dans laquelle un homme parle est un monde dans lequel il vit et s'efforce.* »

Plus récemment, le système de génération de texte GPT-3.5, annoncé comme ses prédécesseurs comme un apprenant multitâche (Radford et al., 2019) produit la traduction « *La langue qu'elle parle est celle d'une personne qui a en elle un monde où elle vit et lutte.* ». Si cette traduction n'est pas exacte, elle est nettement plus proche de la réalité que celle du m2m100 original. En revanche la traduction dans le sens français⟶breton de la même phrase donne « *An teunga a implij ur vro eo ur bed en e ober a blij ha emdroadur* », qui contient un mot inventé (« *teunga* »), est grammaticalement incorrecte et globalement très éloignée d'une traduction correcte. Des requêtes subséquentes aboutissent à d'autres traductions, toutes erronées, mais toutes délivrées par le modèle avec une certitude absolue.

Cette technologie sûre d'elle-même est reprise et promue par d'autres. Ainsi, un non-locuteur du breton diffuse depuis 2023 un chat construit sur ce même modèle de génération. Le breton diffusé est erratique et propose en page d'accueil de « *kregiñ ar c'hat* », ce qui est supposé signifier « commencer le chat ». En breton réel cependant, cela peut se traduire par « commencer le "gat" », mot étrange qui peut être compris à la limite comme une forme abrégée d'un nom féminin, soit de « *gast* » « putain », soit de l'emprunt à l'acronyme anglophone GATT « Accord général sur les tarifs douaniers et le commerce ». Le reste du site et des productions du chat sont de qualité égale. Les promesses du site n'en sont pas moins cruciales pour une langue minorisée : « Que vous soyez débutant ou avancé, notre chatbot en breton peut vous aider à progresser. », ou bien « Si vous cherchez à apprendre la langue bretonne ou à améliorer votre niveau, vous êtes au bon endroit. ». Les dégâts potentiels sont considérables. Contacté en 2023 pour l'alerter sur les fautes diffusées, l'auteur n'a pas donné suite et a intégré le GPT store en 2024.

2 Est-il possible aux subalternes de parler ?

Comme toute problématique où la dimension de pouvoir est en jeu, l'enjeu central est l'écoute des subalternes (Chakravorty Spivak, 1988), et ce qui est mis en place pour que cette écoute advienne. La communauté internationale ne s'est pas dotée de moyens pour protéger les langues sous-outillées de la publicité mensongère. Or, cette brèche fait des dégâts concrets (pour l'exemple algonquien, voir Junker (2024)). Nous listons ici ces effets, des plus ponctuels aux plus systémiques.

Les outils sont utilisés et promus par des non-locuteurs inconscients de la mauvaise qualité des résultats. Les langues minorisées doivent alors faire face à une nouvelle source de données publiques qui répandent des formes erratiques de la langue. Cela est particulièrement dévastateur pour les langues très dialectalisées, où des locuteurs natifs de dialectes traditionnels peuvent les interpréter comme des formes standard qui leur seraient inconnues, formes à apprendre eux-mêmes et à transmettre.

Ces langues paraissent en surface être raisonnablement outillées, et les politiques linguistiques globales, au niveau des États et des unions d'États, faillissent à appréhender leurs urgences.

Les locuteurs sont mis en conflit d'autorité linguistique avec les outils, les outils de transcription, de synthèse vocale ou de traduction mais aussi et surtout avec les outils de génération de langage, et ne peuvent que subir ce rapport de force inégal. Cela aggrave le sentiment de dépossession et de mise hors puissance qui est par ailleurs caractéristique des locuteurs des langues minorisées.

L'effet pervers le plus systémique est que ces outils vont rester présents dans le contenu du discours publicitaire tant que les communautés parlantes minorisées resteront inaudibles sur la réalité des performances réelles des outils. Conserver cet état de fait est donc dangereux, car il fournit un intérêt durable à des entreprises internationales puissantes de silencier les communautés parlantes qui se trouvent dans des situations sociales fragiles. Laisser le feedback qualitatif à charge des communautés des langues minorisées est une stratégie dont le résultat est prévisible. Cela revient à s'attendre à ce que des locuteurs qui ne sont pas nécessairement bilingues en anglais, rarement universitaires, rarement formés en informatique et linguistique, fassent émerger leurs évaluations sur des plates-formes spécialisées qu'ils auraient identifiés seuls, et lesquelles sont typiquement très peu interactives, dans l'espoir d'être entendus de gens qui y ont un intérêt opposé. Faire ce choix, c'est faire le choix de silencier les communautés parlantes.

Ce qui vaut pour la publicité mensongère des

grandes entreprises mondiales vaut par parité d'argument pour la recherche universitaire. Les approches quantitatives qui excluent des stratégies de remontées de validation de la qualité de la part des locuteurs eux-mêmes ne peuvent être valorisées que dans l'exacte mesure où cet état de fait persiste. Conserver cet état de fait implique de donner aux chercheurs développant des approches principalement quantitatives des motivations systémiques durables pour écarter, refuser ou sous-financer des approches incluant des validations qualitatives. Cet effet est d'autant plus redoutable qu'il agit sur le temps long, en sélectionnant à bas bruit la recherche de demain en double aveugle, dans les jurys, les évaluations d'articles et de conférences et les instances de création et de fléchage de postes. Cet effet est de plus aggravé dans le domaine scientifique en ces périodes de refonte autour des avancées de l'« intelligence artificielle », car les fondations d'aujourd'hui décident de nos impossibilités de demain.

3 Recommandations

Nous recommandons de développer des outils de diffusion des données langagières qui incluent des fonctionnalités d'évaluation des données. Il doit être aisé pour des experts de langue et des linguistes, avec un coût d'entrée technique moindre, de commenter sur la qualité des paquets de données rendues disponibles pour les développeuses et développeurs.

Une telle évaluation n'est cependant qu'un pis-aller tant qu'elle reste externe au développement de données et d'outils, et ne peut être qu'un complément à une réelle intégration des linguistes et des experts de langues à chaque étape de ces processus de développement, et non pas uniquement dans des rôles de consultants ou de subalternes.

Références

Chakravorty Spivak, Gayatri (1988). « Can the Subaltern Speak ». In : *Marxism and the Interpretation of Culture*. University of Illinois Press. URL : https://jan.ucc.nau.edu/~sj6/Spivak%20CanTheSubalternSpeak.pdf.

Fan, Angela et al. (2021). « Beyond English-Centric Multilingual Machine Translation ». In : *The Journal of Machine Learning Research* 22.1 (1er jan. 2021), 107 :4839-107 :4886. URL : https://dl.acm.org/doi/abs/10.5555/3546258.3546365.

Jouitteau, Mélanie (2009-2024). *ARBRES, Wikigrammaire Des Dialectes Du Breton et Centre de Ressources Pour Son Étude Linguistique Formelle*. URL : http://arbres.iker.cnrs.fr.

Junker, Marie-Odile (2024). « Data-Mining and Extraction : The Gold Rush of AI on Indigenous Languages ». In : *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*. St. Julians, Malta : Association for Computational Linguistics, mars 2024, p. 52-57. URL : https://aclanthology.org/2024.computel-1.8.

Post, Matt (2018). « A Call for Clarity in Reporting BLEU Scores ». In : *Proceedings of the Third Conference on Machine Translation : Research Papers*. WMT 2018. Brussels, Belgium : Association for Computational Linguistics, oct. 2018, p. 186-191. DOI : 10.18653/v1/W18-6319. URL : https://aclanthology.org/W18-6319.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei et Ilya Sutskever (2019). *Language Models Are Unsupervised Multitask Learners*. preprint.

Tyers, Francis M. (2009). « Rule-Based Augmentation of Training Data in Breton-French Statistical Machine Translation ». In : *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*. EAMT 2009 (Barcelona, España). European Association for Machine Translation, 14 mai 2009. URL : https://aclanthology.org/2009.eamt-1.29.