En 2011, aux États-Unis, l'ordinateur Watson, d'IBM, triomphe au jeu Jeopardy. Nourri avec l'équivalent de 1 million de livres, il capte les questions posées en langage naturel par le présentateur, le sens des mots, mais aussi s'il faut ou non répondre à une question. Une performance pas si élémentaire…

En 2019, ils ont créé WinoGrande, un lot bien plus étoffé de schémas Winograd : il contient 44 000 phrases, contre quelques centaines jusque-là. Pour obtenir autant d'exemples, les chercheurs ont fait appel à Amazon Mechanical Turk, une plateforme pour proposer à des humains, contre rémunération, des tâches plus ou moins complexes. Il a été demandé à chaque travailleur d'écrire quelques paires de phrases, avec des contraintes pour s'assurer que l'ensemble aborderait divers sujets, mais avec la possibilité que ces paires puissent à présent différer de plus d'un mot.

Les chercheurs ont ensuite tenté d'éliminer les phrases propices à des raccourcis statistiques, en soumettant chacune d'entre elles à une IA assez peu sophistiquée et en écartant toutes celles qui étaient trop simples à résoudre. Comme ils s'y attendaient, les phrases qui restaient offraient un défi bien plus difficile pour les machines que la collection originale. Tandis que les humains continuaient d'obtenir des scores très élevés, les modèles de langage en réseau neuronal qui égalaient la performance humaine sur le lot d'origine ont vu leur score s'effondrer avec le lot WinoGrande. Ce nouveau défi restaurait le statut de « test de compréhension de sens commun » pour les schémas Winograd – du moment que les phrases étaient scrupuleusement filtrées pour assurer leur imperméabilité à des recherches Google.

> Il est souvent difficile de savoir si des systèmes d'IA comprennent véritablement le langage, ou d'autres données, qu'ils traitent

Cependant, une autre surprise allait surgir. Dans les deux années qui suivirent la publication de la collection WinoGrande, les modèles de langage en réseau neuronal ont continué de croître, et plus ils sont larges, meilleur semble être leur score à ce nouveau défi. Fin 2021, les meilleurs programmes actuels – d'abord entraînés avec des térabits de textes, puis avec des milliers d'exemples WinoGrande – obtenaient

près de 90 % de bonnes réponses (contre 94 % pour les humains). Cette hausse de performance est presque entièrement imputable à l'augmentation de la taille des modèles de langage en réseau neuronal et de la quantité de leurs données d'entraînement.

Ces réseaux encore plus larges ont-ils enfin atteint un niveau de compréhension similaire au nôtre ? Encore une fois, c'est peu probable. Les résultats de WinoGrande s'accompagnent d'importantes mises en garde. Par exemple, parce que les phrases dépendent des travailleurs d'Amazon Mechanical Turk, la qualité et la cohérence de l'écriture sont assez inégales. L'IA utilisée pour filtrer les phrases « non imperméables à Google » peut avoir été trop peu sophistiquée pour repérer tous les potentiels raccourcis statistiques qu'un énorme réseau neuronal pourrait emprunter, et elle ne s'appliquait qu'à des phrases individuelles, si bien que certaines des phrases qui restaient ont fini par perdre leur « jumelle ». Une étude postérieure à ces travaux a montré que les modèles de langage en réseau neuronal testés uniquement avec des phrases jumelles, et devant répondre correctement aux deux, sont bien moins précis que les humains, ce qui indique que le résultat de 90 % vu plus tôt est moins significatif qu'il ne pouvait le paraître.

Au final, que retenir de cette saga Winograd ? La leçon principale est qu'il est souvent difficile de déterminer, à partir de leur performance lors d'un défi donné, si des systèmes d'IA comprennent véritablement le langage (ou d'autres données) qu'ils traitent. Nous savons que les réseaux neuronaux utilisent souvent des raccourcis statistiques – au lieu de vraiment faire preuve d'une compréhension semblable à celle des humains – pour obtenir de bonnes performances sur les schémas Winograd et sur d'autres bancs d'essais orientés vers une « compréhension générale du langage ».

Le nœud du problème, à mon avis, est que comprendre le langage nécessite de comprendre le monde, et notamment ce que signifie que « la voiture de sport a doublé la camionnette de la poste parce qu'elle était plus lente ». Cela suppose de savoir ce que sont des voitures de sport et des camionnettes de la poste, que des voitures peuvent se « doubler », et, à un niveau encore plus fondamental, que les véhicules sont des objets qui existent et interagissent dans le monde, conduits par des humains avec leurs propres objectifs.

Ces connaissances, nous, humains, les tenons pour acquises, mais ce n'est pas le cas des machines. Et il est peu probable que ce soit écrit explicitement dans le texte d'entraînement de n'importe quel modèle de langage. Certains chercheurs en cognition estiment que les humains, pour apprendre le langage, s'appuient sur un noyau de connaissances, prélinguistiques et innées, de l'espace, du temps et de nombreuses autres propriétés essentielles du monde. Si nous voulons que les machines maîtrisent le langage comme nous, nous devrons d'abord les doter des principes primordiaux avec lesquels nous naissons. Et pour évaluer leur niveau de compréhension, nous devrions commencer par évaluer leur capacité à saisir ces principes, ce qu'on pourrait appeler une « métaphysique infantile ».

Entraîner et évaluer des machines au niveau d'intelligence d'un nourrisson peut apparaître comme un pas de géant en arrière par rapport aux prouesses de Watson et de GPT-3. Mais si l'objectif est une compréhension authentique et digne de confiance, il se peut que ce soit le seul chemin vers des machines vraiment capables de comprendre à quoi « il » ou « elle » fait référence dans une phrase.

— L'autrice —

> Melanie Mitchell est professeuse de complexité à l'institut de Sante Fe, au Nouveau-Mexique, aux États-Unis.

— À lire —

> Y. Elazar *et al.*, Back to square one : Artifact detection, training and common sense disentanglement in the Winograd schema, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

> B. Taira *et al.*, A pragmatic assessment of Google Translate for emergency department instructions, *Journal of General Internal Medicine*, 2021.

> P. Amsili et O. Seminck, Schémas Winograd en français : une étude statistique et comportementale, *Actes de TALN 2017*, 2017.

> T. Brown *et al.*, Language models are few-shot learners, https://arxiv.org/abs/2005.14165.