En bref

> Donner aux machines la maîtrise du langage naturel est, de longue date, un objectif de l'intelligence artificielle.

> Entraînés de façon adéquate, les réseaux de neurones comme GPT-3 semblent capables de performances spectaculaires. Mais comprennent-ils vraiment ?

> Du test de Turing aux schémas Winograd, il n'est pas aisé d'établir un test prouvant une vraie compréhension, notamment parce que les programmes recourent à des raccourcis statistiques.

> Les machines ne sauront sans doute comprendre le langage que lorsqu'elles sauront aussi comprendre le monde qui les entoure.

de « test de Turing ». Une machine et un humain, tous deux cachés, sont en compétition pour convaincre un juge humain de leur humanité en n'usant que de la conversation écrite. Si le juge est incapable de trancher, pensait Turing, nous devrions alors considérer que la machine pense – et, en effet, comprend.

Malheureusement, Turing a sous-estimé la propension des humains à se laisser duper. Même de simples robots de conversation, tels qu'Eliza – l'ersatz de psychothérapeute conçu par Joseph Weizenbaum en 1960 –, ont réussi à faire croire à des testeurs qu'ils discutaient avec un être doué de compréhension, alors même qu'ils savaient parler à une machine.

MAÎTRE DU MONDE POUR UN PRONOM

Dans un article de 2012, les informaticiens Hector Levesque, Ernest Davis et Leora Morgenstern ont proposé un test plus objectif, qu'ils ont baptisé « challenge du schéma Winograd ». Ce test a, depuis, été adopté par la communauté du langage de l'IA comme moyen, peut-être le meilleur, d'évaluer le degré de compréhension d'une machine – bien que, nous le verrons, il ne soit pas parfait. Un schéma Winograd, nommé d'après le chercheur en linguistique Terry Winograd, de l'université Stanford, consiste en une paire de phrases différant d'un mot exactement, chacune suivie d'une question. Voici deux exemples :

Paire 1, phrase 1 : J'ai versé de l'eau de la bouteille dans la tasse jusqu'à ce qu'elle soit *pleine*.

Question : Qu'est-ce qui était rempli, la bouteille ou la tasse ?

Cet article a d'abord été publié en anglais par ***Quanta Magazine,*** *une publication en ligne indépendante, soutenue par la Simons Foundation afin de favoriser la diffusion des sciences : bit.ly/3MFPKtr*

Paire 1, phrase 2 : J'ai versé de l'eau de la bouteille dans la tasse jusqu'à ce qu'elle soit *vide*.

Question : Qu'est-ce qui était vide, la bouteille ou la tasse ?

Paire 2, phrase 1 : L'oncle de Joe peut encore le battre au tennis, bien qu'il soit trente ans plus *vieux*.

Question : Qui est plus vieux, Joe ou l'oncle de Joe ?

Paire 2, phrase 2 : L'oncle de Joe peut encore le battre au tennis, bien qu'il soit trente ans plus *jeune*.

Question : Qui est plus jeune, Joe ou l'oncle de Joe ?

Dans chaque paire, la différence d'un mot *(en italique)* peut changer la personne ou la chose à laquelle le pronom fait référence. Répondre correctement à ces questions semble nécessiter une compréhension de type sens commun. Les schémas Winograd sont conçus précisément pour évaluer cela, en réduisant la vulnérabilité du test de Turing aux juges humains non fiables et aux astuces des robots conversationnels. En particulier, les auteurs ont conçu quelques centaines de schémas qu'ils pensent être imperméables à Google : une machine ne devrait pas être capable de lancer une recherche Google (ou tout autre moteur de recherche) pour répondre à ces questions correctement.

Ces schémas ont été l'objet d'une compétition en 2016 durant laquelle le programme vainqueur a répondu juste à seulement 58 % des phrases – un résultat à peine meilleur que celui qu'elle aurait obtenu en devinant les réponses au hasard. Ce qui fait dire malicieusement à Oren Etzioni, éminent chercheur en IA : « Quand

une intelligence artificielle ne peut déterminer à qui un pronom fait référence dans une phrase, il est difficile d'imaginer qu'elle puisse prendre un jour le contrôle du monde.»

Cependant, la capacité des programmes d'IA à résoudre les schémas Winograd a grimpé en flèche avec l'avènement des modèles de langage en réseau neuronal. En 2020, une étude d'OpenAI relatait que GPT-3, qui relève de cette catégorie d'IA, obtenait de bonnes réponses sur près de 90 % des phrases dans un lot de référence de phrases de ce type. D'autres modèles de langage sont encore plus performants après s'être entraînés spécifiquement à ces tâches. Fin 2021, les modèles de langage en réseau neuronal atteignaient 97 % de précision sur un lot particulier de schémas Winograd qui font partie d'une compétition de compréhension des langues par les IA, nommée SuperGlue. C'est à peu près équivalent aux performances humaines. Cela signifie-t-il que les modèles de langage en réseau neuronal ont atteint un niveau de compréhension comparable au nôtre ?

ÊTRE COLLÉ À LA COMPÉTITION SUPERGLUE

Pas forcément. Malgré tous les efforts de leurs créateurs, ces schémas Winograd n'étaient en fait pas imperméables à Google. Ces défis, comme beaucoup d'autres tests actuels de compréhension du langage par les IA, recourent parfois à des raccourcis statistiques grâce auxquels les réseaux neuronaux sont performants sans comprendre. Prenez, par exemple, les phrases «la voiture de sport a doublé la camionnette de la poste, car elle était plus rapide» et «la voiture de sport a doublé la camionnette de la poste, car elle était plus lente». Un modèle de langage entraîné avec un large corpus de phrases en français aura assimilé la corrélation entre «voiture de sport» et «rapide», et entre «camionnette de la poste» et «lent»; il peut donc répondre correctement uniquement grâce à ces corrélations. De fait, beaucoup de schémas Winograd de la compétition SuperGlue autorisent ce genre de lien statistique.

> Turing a sous-estimé la propension des humains à se laisser duper. Même de simples robots de conversation peuvent faire illusion

Plutôt qu'abandonner ces schémas en tant que test de compréhension, un groupe de chercheurs de l'institut Allen (du nom d'un cofondateur de Microsoft) pour l'IA, à Seattle, a essayé de régler certains de leurs problèmes.

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