THE CONVERSATION

L'expertise universitaire, l'exigence journalistique

Une des techniques utilisées par les systèmes d'intelligence artificielle exploite les statistique en grandes dimensions pour relier des mots entre eux, ou des gens partageant des traits de caractères, et les associer. vi73, Shutterstock

De Cambridge Analytica à ChatGPT, comprendre comment l'IA donne un sens aux mots

Publié: 19 mai 2023, 12:11 CEST

Frederic Alexandre

Directeur de recherche en neurosciences computationnelles, Université de Bordeaux, Inria

Un des problèmes que l'IA n'a toujours pas résolu aujourd'hui est d'associer des symboles – des mots par exemple – à leur signification, ancrée dans le monde réel – un problème appelé l'« ancrage du symbole ».

Par exemple, si je dis : « le chat dort sur son coussin car il est fatigué », la plupart des êtres humains comprendra sans effort que « il » renvoie à « chat » et pas à « coussin ». C'est ce qu'on appelle un raisonnement de bon sens.

En revanche, comment faire faire cette analyse à une IA ? La technique dite de « plongement lexical », si elle ne résout pas tout le problème, propose cependant une solution d'une redoutable efficacité. Il est important de connaître les principes de cette technique, car c'est celle qui est utilisée dans la plupart des modèles d'IA récents, dont ChatGPT… et elle est similaire aux techniques utilisées par Cambridge Analytica par exemple.

Le plongement lexical, ou comment les systèmes d'intelligence artificielle associent des mots proches

Cette technique consiste à remplacer un mot (qui peut être vu comme un symbole abstrait, impossible à relier directement à sa signification) par un vecteur numérique (une liste de nombres). Notons que ce passage au numérique fait que cette représentation peut être directement utilisée par des réseaux de neurones et bénéficier de leurs capacités d'apprentissage.

Plus spécifiquement, ces réseaux de neurones vont, à partir de très grands corpus de textes, apprendre à plonger un mot dans un espace numérique de grande dimension (typiquement 300) où chaque dimension calcule la probabilité d'occurrence de ce mot dans certains contextes. En simplifiant, on remplace par exemple la représentation symbolique du mot « chat » par 300 nombres représentant la probabilité de trouver ce mot dans 300 types de contextes différents (texte historique, texte animalier, texte technologique, etc.) ou de co-occurrence avec d'autres mots (*oreilles*, *moustache* ou *avion*).

Plonger dans un océan de mots et repérer ceux qui sont utilisés conjointement, voilà une des phases de l'apprentissage pour ChatGPT. Amy Lister/Unsplash, CC BY

Même si cette approche peut sembler très pauvre, elle a pourtant un intérêt majeur en grande dimension : elle code des mots dont le sens est proche avec des valeurs numériques proches. Ceci permet de définir des notions de proximité et de distance pour comparer le sens de symboles, ce qui est un premier pas vers leur compréhension.

Pour donner une intuition de la puissance de telles techniques (en fait, de la puissance des statistiques en grande dimension), prenons un exemple dont on a beaucoup entendu parler.

Relier les traits psychologiques des internautes à leurs « likes » grâce aux statistiques en grande dimension

C'est en effet avec une approche similaire que des sociétés comme Cambridge Analytica ont pu agir sur le déroulement d'élections en apprenant à associer des préférences électorales (représentations symboliques) à différents contextes d'usages numériques (statistiques obtenues à partir de pages Facebook d'usagers).

Leurs méthodes reposent sur une publication scientifique parue en 2014 dans la revue PNAS, qui comparait des jugements humains et des jugements issus de statistiques sur des profils Facebook.

L'expérimentation reportée dans cette publication demandait à des participants de définir certains de leurs traits psychologiques (sont-ils consciencieux, extravertis, etc.), leur donnant ainsi des étiquettes symboliques. On pouvait également les représenter par des étiquettes numériques comptant les « likes » qu'ils avaient mis sur Facebook sur différents thèmes (sports, loisirs, cinéma, cuisine, etc.). On pouvait alors, par des statistiques dans cet espace numérique de grande dimension, apprendre à associer certains endroits de cet espace à certains traits psychologiques.

Ensuite, pour un nouveau sujet, uniquement en regardant son profil Facebook, on pouvait voir dans quelle partie de cet espace il se trouvait et donc de quels types de traits psychologiques il est le plus proche. On pouvait également comparer cette prédiction à ce que ses proches connaissent de ce sujet.

Le résultat principal de cette publication est que, si on s'en donne les moyens (dans un espace d'assez grande dimension, avec assez de « likes » à récolter, et avec assez d'exemples, ici plus de 70000 sujets), le jugement statistique peut être plus précis que le jugement humain. Avec 10 « likes », on en sait plus sur vous que votre collègue de bureau ; 70 « likes » que vos amis ; 275 « likes » que votre conjoint.

Être conscients de ce que nos « likes » disent sur nous

Cette publication nous alerte sur le fait que, quand on recoupe différents indicateurs en grand nombre, nous sommes très prévisibles et qu'il faut donc faire attention quand on laisse des traces sur les réseaux sociaux, car ils peuvent nous faire des recommandations ou des publicités ciblées avec une très grande efficacité. L'exploitation de telles techniques est d'ailleurs la principale source de revenus de nombreux acteurs sur Internet.

likes peints sur un mur argenté

Nos likes et autres réaction sur les réseaux sociaux en disent beaucoup sur nous, et ces informations peuvent être exploitées à des fins publicitaires ou pour des campagnes d'influence. George Pagan III/Unsplash, CC BY

Cambridge Analytica est allée un cran plus loin en subtilisant les profils Facebook de millions d'Américains et en apprenant à associer leurs « likes » avec leurs préférences électorales, afin de mieux cibler des campagnes électorales américaines. De telles techniques ont également été utilisées lors du vote sur le Brexit, ce qui a confirmé leur efficacité.

Notons que c'est uniquement l'aspiration illégale des profils Facebook qui a été reprochée par la justice, ce qui doit continuer à nous rendre méfiants quant aux traces qu'on laisse sur Internet.

Calculer avec des mots en prenant en compte leur signification

En exploitant ce même pouvoir des statistiques en grande dimension, les techniques de plongement lexical utilisent de grands corpus de textes disponibles sur Internet (Wikipédia, livres numérisés, réseaux sociaux) pour associer des mots avec leur probabilité d'occurrence dans différents contextes, c'est-à-dire dans différents types de textes. Comme on l'a vu plus haut, ceci permet de considérer une proximité dans cet espace de grande dimension comme une similarité sémantique et donc de calculer avec des mots en prenant en compte leur signification.

Un exemple classique qui est rapporté est de prendre un vecteur numérique représentant le mot *roi*, de lui soustraire le vecteur (de même taille car reportant les probabilités d'occurrence sur les mêmes critères) représentant le mot *homme*, de lui ajouter le vecteur représentant le mot *femme*, pour obtenir un vecteur très proche de celui représentant le mot *reine*. Autrement dit, on a bien réussi à apprendre une relation sémantique de type « A est à B ce que C est à D ».

Le principe retenu ici pour définir une sémantique est que deux mots proches sont utilisés dans de mêmes contextes : on parle de « sémantique distributionnelle ». C'est ce principe de codage des mots qu'utilise ChatGPT, auquel il ajoute d'autres techniques.

Ce codage lui permet souvent d'utiliser des mots de façon pertinente ; il l'entraîne aussi parfois vers des erreurs grossières qu'on appelle hallucinations, où il semble inventer des nouveaux faits. C'est le cas par exemple quand on l'interroge sur la manière de différencier des œufs de poule des œufs de vache et qu'il répond que ces derniers sont plus gros. Mais est-ce vraiment surprenant quand on sait comment il code le sens des symboles qu'il manipule ?

Sous cet angle, il répond bien à la question qu'on lui pose, tout comme il pourra nous dire, si on lui demande, que les vaches sont des mammifères et ne pondent pas d'œuf. Le seul problème est que, bluffés par la qualité de ses conversations, nous pensons qu'il a un raisonnement de bon sens similaire au nôtre : qu'il « comprend » comme nous, alors que ce qu'il comprend est juste issu de ces statistiques en grande dimension.