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

Frédéric Alexandre

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

ALEXANDRE Frédéric, Centre Inria de l'université de Bordeaux, CNRS, Bordeaux INP

Un des problèmes majeurs que l'IA n'a toujours pas résolus aujourd'hui est celui de l'ancrage du symbole (*symbol grounding problem*), c'est-à-dire comment des symboles (des mots par exemple) peuvent être associés à leur signification. 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 le raisonnement de bon sens. Par contre, ce raisonnement va être beaucoup plus difficile en IA et je vais expliquer pourquoi. Mais je vais aussi présenter la technique dite de [plongement lexical](https://fr.wikipedia.org/wiki/Plongement_lexical) qui, si elle ne résout pas tous les problèmes, est cependant d'une redoutable efficacité. Il est important de connaitre les principes de cette technique car c'est celle qui est utilisée dans la plupart des modèles d'IA récents, dont ChatGPT.

Le problème de l'ancrage du symbole a été identifié bien avant l'IA, par des philosophes et des linguistes, et a été résumé, au début du XXème par trois concepts regroupés dans un triangle dit sémiotique, dont les trois sommets renvoient respectivement au symbole (le signifiant), l'objet qu'il représente et le concept (ou le signifié), correspondant à l'image mentale que nous nous sommes construite, suite à nos expériences avec cet objet. Ce sont ces expériences (et en particulier leurs valeurs émotionnelles) et les propriétés de cet objet qu'elles nous ont permis d'extraire qui nous permettent d'avoir ce bon sens pour raisonner à son sujet et qui sont singulièrement absentes d'un système d'IA et qui lui donnent cette limitation d'accès à la sémantique des symboles.

John Searle a illustré ce problème en 1980 avec son expérience de pensée dite de la chambre chinoise. Dans cette chambre, un être humain qui ne parle pas le chinois reçoit des questions écrites en chinois. Il a accès à un ensemble de règles qui lui dit mécaniquement comment réagir à chaque caractère (ou combinaison de caractères) possibles, ce qui lui permet d'écrire une réponse à la question et de l'envoyer à l'extérieur de la chambre. Ceci permet de montrer que si, pour une tâche donnée (un ensemble de questions), on est capable d'écrire l'ensemble des règles qui régissent son traitement, alors un système mécanique (ou un humain qui ne comprend pas le chinois) peut donner l'illusion à l'extérieur qu'il est intelligent. Mais même dans ce cas (et on peut imaginer qu'écrire un tel ensemble de règles n'est pas simple voire impossible pour de nombreuses tâches), l'agent ne comprendra pas ce qu'il fait et n'aura donc pas accès à la signification des symboles qu'il manipule (ce que fera très naturellement toute personne parlant le chinois). Searle en déduisait alors qu'une approche purement mécanique, comme celle d'un ordinateur, quelle que soit la complexité des règles ou des algorithmes qu'il manipule, ne donnera donc jamais accès à la capacité de comprendre le sens de ce que l'on fait.

Considérant que notre cerveau est tout à la fois capable de manipuler des symboles (des mots par exemple) et d'apprendre leur signification à travers ses interactions avec le monde, Stevan Harnard a repris ces idées en 1990 pour reformuler le problème de l'ancrage du symbole, en proposant qu'on pouvait le résoudre à une condition : L'interprétation sémantique d'un système de symboles doit être faite de façon intrinsèque à ce système, si on

veut qu'il ait effectivement accès au sens des symboles qu'il manipule. Il s'agirait donc d'ancrer notre système de symboles dans nos expériences sensorimotrices avec le monde, pour apprendre des relations symboles-significations. De cette conception est née l'approche de l'IA incarnée, qui postule qu'un tel système (un robot par exemple) devrait avoir un corps pour ressentir et interagir avec le monde, si il veut pouvoir donner un sens aux symboles qu'il manipule. Il reste maintenant à voir comment faire ça en pratique. C'est ce que propose, dans le cadre très restreint de l'analyse de textes, le plongement lexical.

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. En particulier, 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). Même si cette approche peut sembler très pauvre, elle a pourtant un intérêt majeur en grande dimension : elle code avec des valeurs numériques proches des mots dont le sens est proche, ce qui va alors permettre 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 pour des phénomènes avec des régularités telles qu'on en voit dans notre monde cognitif), prenons un exemple similaire dont beaucoup ont entendu parler.

C'est en effet avec une approche similaire que des sociétés comme [Cambridge Analytica](https://fr.wikipedia.org/wiki/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 subtilisées 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 quelques dizaines de milliers de participants de définir certains de leurs traits psychologiques (sont-ils consciencieux, extravertis, etc.). Ces participants avaient donc des étiquettes (dites symboliques) représentant ces traits. On pouvait également les représenter par une étiquette (dite numérique) comptant les 'Likes' qu'ils avaient mis sur Facebook sur différents sujets (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 connait de ce sujet ses proches. Le résultat principal de cette publication est que, si on s'en donne les moyens (dans un espace d'assez grande dimension, donc avec assez de 'Likes' à récolter, et avec assez d'exemples, ici plus de 70 000 sujets), le jugement statistique peut être plus précis que le jugement humain. Autrement dit,

qu'avec 10 Likes, on en sait plus sur vous que votre collègue de bureau ; 70 Likes que vos amis ; 275 Likes que votre conjoint. Cette publication avait tout d'abord comme but de nous alerter 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 de nombreux acteurs sur internet 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 revenu de ces acteurs.

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 pour mieux cibler des campagnes électorales. 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 que l'on laisse sur Internet. En exploitant ce même pouvoir des statistiques en grand dimension, les techniques de plongement lexical utilisent de grands corpus de textes (que l'on trouve facilement sur Internet (Wikipédia, livres numérisés, réseaux sociaux) pour associer des mots avec leur probabilité d'occurrence dans différents contextes (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 le vecteur numérique représentant le mot Roi, de lui soustraire le vecteur (de même taille car reportant les probabilité 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 (dans ce monde lexical) pour définir la sémantique est que deux mots proches sont utilisés dans des mêmes contextes. On parle ici de sémantique distributionnelle. C'est ce principe de codage des mots qu'utile ChatGPT, auquel il ajoute d'autres techniques, comme on peut le voir dans le [texte] (« De quoi ChatGPT est-il le nom].

On a indiqué plus haut que coder avec des valeurs numériques proches des mots dont le sens est proche permet d'aider à leur compréhension. On peut se demander ce qu'il manque pour se rapprocher encore de la façon qu'a notre cerveau de comprendre un concept. On aborde cette question dans le [texte] (ChatGPT est-il intelligent comme un humain ?).