Extraire une mine d'or d'informations de forte technicité grâce à l'intelligence artificielle, c'est désormais possible !

Par Charles Teissèdre, Directeur de la R&D – Synapse Développement & Thiziri Belkacem, ingénieure de recherche – Synapse Développement.

Certaines industries et organisations gèrent des connaissances métier particulièrement pointues (aéronautique, défense, nucléaire, pharmacologie, spatial…). Le langage propre à ces métiers est riche en termes et expressions peu communes. Conséquence directe : l'accès à l'information dans leurs corpus de textes techniques est particulièrement complexe à mettre en place avec des méthodes classiques. Désormais, certaines technologies basées sur l'intelligence artificielle relèvent haut la main ce défi du partage des connaissances, voici comment.

Connaissez-vous les définitions de Gyro compas[1] , blimp[2], spallation[3] ou encore alcaloïde[4] ? Non ? Vous êtes sûr ? C'est bien normal, on vous rassure ! Ces termes, propres à l'aviation, au nucléaire ou à la pharmacologie, ne sont clairement pas du langage courant. Pourtant, leur importance est capitale dans ces industries. En effet, ces dernières (et bien d'autres), utilisent en permanence des corpus de textes à très forte technicité. Dans ce contexte, comment rendre possible l'accès à l'information au sein de ces entités sans supervision permanente d'un expert du sujet ? Car évidemment, il est impensable de mobiliser des équipes d'ingénieurs et de docteurs en telle ou telle spécialité pour la « simple » gestion documentaire d'une société, même hautement spécialisée. Une nouvelle approche de traitement automatique du langage relève ce défi de taille.

Le contenu n'est plus la seule source d'apprentissage de l'intelligence artificielle

Pour parvenir à extraire, trier et mettre à disposition des informations précises d'industries et d'organisations hautement spécialisées, il est particulièrement intéressant d'agir sur deux caractéristiques essentielles : le contenu informationnel, bien sûr, mais aussi, et c'est là que réside toute la nouveauté, la structure organisationnelle des documents, ainsi que la syntaxe et différents aspects liés à la langue elle-même. En effet, ces informations sont de formidables leviers d'apprentissage pour un modèle intelligent.

Chacun le comprend aisément, la structure d'un document traduit une hiérarchisation et une organisation des informations contenues dans ce dernier, selon une certaine logique. C'est cette logique que l'intelligence artificielle analyse, en plus du contenu lui-même. En outre, les aspects syntaxiques et sémantiques liés à chaque langue jouent également un rôle important. Car le langage rédactionnel diffère d'un domaine à un autre et devient plus spécifique lorsqu'il s'agit d'articles ou de documents scientifiques destinés à un public bien ciblé. Ceci est d'autant plus vrai dans les domaines très spécialisés, dans lesquels les ressources documentaires sont particulièrement normées.

Jusqu'ici, les modèles d'intelligence artificielle pré-entraînés sur des corpus de langue dits « tout domaine » peinaient à analyser un vocabulaire et une langue régis par des contraintes différentes de celles des langues naturelles et propres à un métier ou parfois même à une organisation. Si ces modèles éprouvent des difficultés, c'est en partie parce qu'ils ne prennent pas en compte la structure documentaire, ni les spécificités de chaque langue et encore moins la syntaxe. Pour être efficace, l'extraction d'informations dans ce type de textes doit donc se faire par une approche différente, basée désormais sur la forme en plus du fond.

Comment fonctionne cette nouvelle brique technologique ?

Comme toute forme d'intelligence (artificielle ou non), un entraînement est indispensable. Pour « muscler » ces modèles intelligents, une supervision est nécessaire. Mais contrairement aux approches plus classiques, dans le cas qui nous intéresse, cette dernière est réduite au minimum. Elle ne mobilise donc pas toute une batterie d'experts sur le long terme ! Cette supervision consiste simplement en un cycle itératif de validation des extractions dans un premier temps, avant que le modèle exécute lui-même ses propres entrainements.

Très concrètement, cela signifie que la machine propose des exemples d'extractions à un expert qui les valide ou les invalide. Ainsi, l'intelligence artificielle apprend très vite à faire des prédictions de plus en plus précises, même sur des documents au langage hautement technique. Vous l'aurez compris, cette performance est rendue possible par la compréhension de l'organisation logique des textes, du langage spécifique de chaque société (et même le jargon interne) en plus de l'analyse (plus classique) du contenu.

Cette approche novatrice évite donc des allers-retours incessants entre des experts métiers d'un côté (qui connaissent la documentation et son contenu), et des ingénieurs en mesure de développer des solutions d'intelligence artificielle pour en extraire des informations. L'entreprise ou l'organisation, dont les données documentaires sont souvent particulièrement sensibles, peut donc se charger seule de la gestion et du partage de l'information une fois la technologie déployée.

Gyro compas[1] : Dans l'univers aéronautique, le Gyro compas est un appareil gyroscopique doté d'un degré de liberté qui permet de conserver une référence de cap de façon beaucoup plus précise qu'un compas magnétique. Il est asservi à une vanne de flux qui permet de le recaler automatiquement en fonction du champ magnétique terrestre. Il est aussi appelé « plateau de route ».

Blimp[2] : Toujours en aéronautique, le mot blimp désigne un dirigeable à enveloppe souple. Le terme blimp serait onomatopéique, c'est-à-dire qu'il symbolise le son que fait l'aéronef quand on frappe l'enveloppe du ballon avec un doigt.

Spallation[3] : La spallation désigne la réaction nucléaire provoquée par des particules accélérées avec une si grande énergie que le noyau atomique qu'elles bombardent « éclate » en éjectant des particules plus légères.

Alcaloïde[4] : Les alcaloïdes désignent des molécules à bases azotées, très majoritairement d'origine végétale. Ce mot est particulièrement utilisé en pharmacologie puisque les alcaloïdes ont permis

d’ouvrir le domaine de la médication dit chimique à partir de la fin du XIXè siècle.