Acclimatation de l'IA à l'Abes : *la période des semis*

Disposant de données riches et structurées, l'Abes a vocation à accueillir l'intelligence artificielle dans ses outils. Plusieurs initiatives encourageantes sont déjà en cours.

L'Abes et ses réseaux produisent ou agrègent beaucoup de données, souvent riches et bien structurées, qu'il s'agit d'exploiter pour offrir des services. Depuis quelques années déjà, il ne fait plus de doute que les techniques de l'intelligence artificielle relèvent de l'état de l'art, et plus seulement de la recherche scientifique. À ce titre, elles doivent rejoindre la boîte à outils de l'Abes, en répondant à de grands axes de réflexion : les chantiers IA ont vocation à s'aligner sur les missions permanentes de l'Abes et les priorités du projet d'établissement, en tenant compte des moyens et compétences dont l'agence dispose et en s'appuyant sur les besoins et les expertises des établissements de ses réseaux, de ses partenaires et de ses homologues à l'international.

L'IA ne remplacera pas le catalogage ni d'autres formes de curation de données. Au contraire, il faut des données de qualité, validées par des professionnels, pour faire apprendre les machines. L'IA peut aider là où le travail humain serait trop fastidieux ou incapable de digérer la masse des documents à décrire et analyser.

IDENTIFIER DES OBJECTIFS PRIORITAIRES ET FAIRE ÉMERGER UNE COMPÉTENCE INTERNE

À partir de ces principes directeurs, de nombreux défis se posent :

Faire émerger une compétence interne, sans recourir exclusivement à l'externalisation. Il faut à la fois essayer de rééquilibrer les spécialisations informatiques actuelles (conception et développement d'applications, gestion d'infrastructures) vers l'ingénierie des données, et compter sur la remarquable capacité des bibliothécaires et des informaticiens de l'Abes à embrasser de nouvelles technologies. Les études du Labo visent à construire avec eux des mises en pratique orientées par des besoins précis, pour que demain on ne parle plus d'IA mais d'outils précis, spécialisés et banalisés.

Identifier les objectifs prioritaires et accessibles

L'étiquette « IA » recouvre différents types de tâches génériques qui ont toutes une application potentielle dans notre contexte : classification automatique (Rameau, Dewey, codes de fonction, types de document, langue), reconnaissance d'entités (repérer une personne dans une mention de responsabilité, un organisme dans une affiliation), liage automatique (d'un nom à IdRef ou Orcid, d'une fonction à un code, d'une manifestation à une œuvre) ; mesure de similarité entre entités (moteur de recommandation, dédoublonnage) ; détection de clusters (repérer des groupes de bibliothèques ou de documents dans la masse des localisations, par exemple). Les possibilités sont immenses, les difficultés inégales.

DEUX ÉTUDES QUI UTILISENT LE *MACHINE LEARNING* POUR AMÉLIORER LES NOTICES BIBLIOGRAPHIQUES SUDOC

Depuis 2021, c'est principalement dans le cadre du Labo, sous forme d'études, que l'Abes fait ses premiers pas.

En 2021 et 2022, le labo de l'Abes a accueilli pendant 6 mois deux étudiants du master IASD (Intelligence artificielle, Systèmes, Données) de l'université de Montpellier, sous la direction du Professeur Pascal Poncelet. Deux études, sur deux thématiques distinctes, ont pu être menées.

La première, effectuée par Min Young Yang, portait sur l'indexation Rameau automatique. Avec l'identification des auteurs, l'indexation automatique est un défi majeur pour la gestion des bases bibliographiques, confrontée à l'inflation de la production scientifique. Dans le cadre de ce stage, il s'agissait de créer un premier prototype qui prédit des concepts Rameau pertinents à partir du titre et du résumé d'un document. Défi redoutable étant donné le caractère subjectif de cette opération intellectuelle et la taille du vocabulaire (100 000 termes). En 2023, l'Abes poursuivra ce travail à travers une collaboration avec une société spécialiste en IA. À terme, si la preuve de concept est concluante, il s'agira d'appliquer l'algorithme aux données de l'Abes mais aussi de le partager sous la forme d'un web service et de code ouvert (éventuellement intégrable dans un outil générique comme Annif[1]).

La seconde étude, menée par Thomas Zaragoza, consistait dans le repérage des auteurs et de leur fonction dans les mentions de responsabilité des notices bibliographiques du Sudoc. Dans plusieurs centaines de milliers de notices du Sudoc, la transcription des mentions de responsabilité n'est pas en accord avec les entrées Auteur : certains auteurs ne sont pas listés séparément dans les zones dédiées, ou bien sans précision de leur rôle. Thomas Zaragoza

[1] https://annif.org

a travaillé à identifier automatiquement dans le texte des mentions de responsabilité les chaînes de caractère qui correspondent à une personne (PER) ou à une fonction (FONCT) :

Hawking / Stephen Finnigan, réalisation ; Stephen Hawking, Stephen Finnigan, Ben Bowie, scénario ; Joe Lovell ; Tina Lovell ; Arthur Pelling [et al.] acteurs

Cette opération de reconnaissance d'entités (NER) a nécessité l'annotation manuelle de milliers de notices, à travers une interface dédiée, pour fournir à l'algorithme d'apprentissage des données de qualité en entrée. Il fallait ensuite aligner les mots évoquant une fonction avec le bon code, ce qui est souvent bien plus difficile que dans l'exemple ci-dessus. L'équipe labo entend approfondir le travail sur ce dernier point.

Yann Nicolas
Responsable du Labo de l'Abes
nicolas@abes.fr

Crédit Adobe stock

QUALINKA : DE L'IA « À L'ANCIENNE » POUR AUTOMATISER LE LIAGE AUX AUTORITÉS

Dès 2012, en participant au projet ANR Qualinka, l'Abes a voulu trouver des solutions techniques pour automatiser le travail de liage et de diagnostic qualité des liens entre notices bibliographiques et notices d'autorité. Les efforts opérationnels ont porté jusqu'à présent sur les liens entre les personnes référencées dans les notices bibliographiques du Sudoc et les notices d'autorité de type personnes physiques du référentiel IdRef. Depuis 2019, le programme Qualinka issu de ces travaux est disponible pour les catalogueurs du Sudoc dans l'application paprika.idref.fr, où il offre une aide à la décision précieuse pour corriger et créer de nouveaux liens.

Un programme d'IA symbolique

Ce courant de l'IA, plus ancien que les méthodes de *machine learning*, repose sur la modélisation des connaissances et des raisonnements humains pour expliciter un ensemble de règles au sein de programmes informatiques capables, au moyen d'approches logiques, de les exécuter pour prendre des décisions. Pour concevoir Qualinka, il a fallu reproduire les différentes étapes qui permettent à un humain de créer des liens bibliographiques. Par exemple, un catalogueur sait d'expérience qu'une personne a tendance à coécrire avec les mêmes personnes ; s'il doit, à partir d'un document, identifier cette personne à une notice d'autorité, il choisira celle dont les documents liés ont les mêmes coauteurs que le document de départ. Évidemment, il n'y a pas toujours de coauteurs et le catalogueur doit, en fait, souvent croiser différentes informations et opérer des pondérations, en particulier pour les cas d'homonymie (plusieurs notices d'autorité avec les mêmes noms et prénoms).

Qualinka prend en entrée un sous-ensemble de références de personnes issues de notices documentaires et de notices d'autorité. Ce sous-ensemble a été préalablement construit à partir d'une recherche sur le nom et le prénom. À chaque référence sont associés des attributs tels que le titre du document, les cocontributeurs, les sujets, la date de publication, les dates de vie, etc. Ces attributs sont comparés au travers de critères prédéfinis, eux-mêmes combinés ensemble dans des règles logiques permettant à Qualinka de décider quelles références correspondent à la même personne.

SudoQual, cadre de développement de nouveaux scénarios d'application

Il n'est pas possible qu'un seul programme puisse traiter tous les problèmes de liage entre divers types d'entités, car les connaissances et les raisonnements impliqués sont différents. En revanche, un cadre technique et méthodologique a été créé et utilisé dans un premier temps pour mettre au point Qualinka. Ce cadre, baptisé SudoQual, permet de réaliser les différentes étapes de configuration d'autres programmes : modélisation, formalisation en règles logiques, élaboration d'algorithmes de décisions, sans oublier l'évaluation du comportement produit et l'ajustement des phases précédentes. SudoQual est d'ores et déjà disponible en *open source*[2]. Cela doit permettre, à l'Abes ou à d'autres, de créer des scénarios d'applications adaptés à différents cas d'usage orientés par un type d'entité, un contexte documentaire, un référentiel cible ou encore une modalité particulière de liage.

Aline Le Provost
Analyste de données bibliographiques, l'Abes
le-provost@abes.fr

[2] https://github.com/abes-esr/sudoqual-framework