IA et approches participatives dans les Humanités Numériques : de la conjonction à la coopération

Olivier Aubert* , Guillaume Raschia*[1], and Benjamin Hervy[2]

[1]Laboratoire d'Informatique de Nantes Atlantique (LINA) – CNRS : UMR6241, Université de Nantes, École Nationale Supérieure des Mines - Nantes – LINA - Faculté des Sciences 2 rue de la Houssinière - BP 92208 44322 NANTES CEDEX 3, France

[2]Maison des Sciences de l'Homme Ange Guépin (MSH Nantes) – MSH Nantes – 5, allée Jacques Berque BP12105 44021 Nantes., France

Résumé

IA et approches participatives dans les Humanités Numériques : de la conjonction à la coopération

Le projet de recherche CIRESFI (*Contrainte et Intégration : pour une réévaluation des spectacles forains et italiens sous l'Ancien Régime*) vise à apporter de nouveaux éclairages sur l'histoire des spectacles. Parmi les angles d'attaque utilisés, le projet RECITAL (*Registres Comptables de la Comédie Italienne*) (Rubellin et Raschi, 2020) développé au sein de CIRESFI vise à exploiter la matière présente dans les registres comptables du théâtre de la Comédie Italienne, en se focalisant particulièrement sur le XVIIIème siècle.

Ainsi, les numérisations sous forme image de 64 registres comptables, couvrant (avec des manques) la période de 1717 à 1794, ont été acquises par le projet. L'enjeu de RECITAL est de procéder à l'extraction des informations présentes dans les 26000 pages des registres, afin de pouvoir les analyser et voir ce que nous enseigne la comptabilité du théâtre sur la structure sociale du public, la popularité du répertoire et des auteurs, les décors et la mise en scène (accessoires), etc.

Le corpus de 26000 pages présente des caractéristiques non homogènes : il est manuscrit, avec plus de 7 rédacteurs différents, donc autant de styles d'écriture. Par ailleurs, la langue utilisée a varié au cours des ans, entre le français et différents types de dialectes italiens. Enfin, la forme et les règles des bilans comptables ont également évolué au cours des années.

Deux approches parallèles ont été testées dès le début du projet : d'une part, une approche purement automatique, mobilisant des algorithmes d'Intelligence Artificielle, visant

*Intervenant

à extraire et transcrire automatiquement les informations. D'autre part, une approche plus manuelle de crowdsourcing, basée sur la mobilisation d'un large public au travers d'une interface participative.

L'approche automatique (A. Granet et al, 2018) s'est retrouvée confrontée aux caractéristiques difficiles du corpus : une forte hétérogénéité des données (scripteurs, langues, formats, etc) ainsi qu'à un manque de données d'entraînement pour l'apprentissage. Elle a permis de faire progresser l'état de l'art, mais n'a pas produit de données utilisables. L'approche de crowdsourcing a pris la forme de la plateforme https://recital.univ-nantes.fr/ utilisant et adaptant le logiciel ScribeAPI. Cette plate-forme offre différentes fonctionnalités précieuses, dont la possibilité d'enchaîner les tâches de marquage, transcription et vérification, ainsi que sa capacité à adapter les formulaires de saisie des informations aux contenus.

La plate-forme permet ainsi de découper le travail en trois étapes : marquage (détermination et indexation des zones présentant des informations à transcrire), transcription (du contenu des zones identifiées dans l'étape de marquage) et enfin vérification (par présentation des contenus saisis).

L'approche participative a permis de générer plus de 115000 transcriptions, qu'il faut ensuite traiter pour homogénéiser les informations ainsi qu'en évaluer la qualité (B. Hervy et al, 2019). En effet, un écueil commun dans les approches participatives est l'évaluation de la qualité des résultats, dont les producteurs ont des compétences et des niveaux de motivation variés. Les données produites font donc l'objet d'une vérification participative au sein de la plate-forme, permettant d'aboutir à une première forme de consensus parmi les transcriptions. Leur évaluation peut ensuite prendre plusieurs formes. On trouve tout d'abord une première évaluation de la cohérence syntaxique et sémantique des données individuelles, tirant parti de la connaissance de leur nature (par exemple, la recette de la journée doit nécessairement être sous une forme numérique), via des scripts python dédiés. On peut également exploiter dans cette phase les informations issues de la phase participative de vérification, qui fournissent une indication de la qualité des données, venant pondérer l'indice de confiance associé à la donnée. Étant donné la nature comptable des informations présentées, on peut également procéder, toujours via des scripts dédiés, à une phase de détection d'incohérences sur un ensemble de données entre elles (recettes mensuelles ne correspondant pas à la somme des recettes quotidiennes par exemple - ce qui peut provenir d'erreurs à la source ou lors de la transcription), mais cette étape nécessite une couverture complète d'un ensemble de données. Enfin, une validation manuelle par les experts du domaine est nécessaire pour évaluer la qualité des informations. Cette phase passe par la mise à disposition d'interfaces permettant une consultation rapide et synthétique de l'ensemble des informations produites. Chacun des traitements mentionnés ci-dessus fait l'objet d'une traçabilité dans le système d'information, de manière à pouvoir retracer la provenance des données nettoyées qui seront finalement présentées aux utilisateurs du système.

Vers une coopération IA/approche participative

Les approches automatiques et participatives ont été dans un premier temps considérées indépendamment. En effet, l'approche participative nécessite un certain temps pour mettre en place l'infrastructure puis récolter un nombre suffisant de données. Nous sommes à présent dans une situation différente, où l'approche participative a déjà permis de couvrir près de

la moitié des pages du corpus. Nous nous attachons donc maintenant à étudier la manière dont on peut combiner les deux approches, en utilisant l'une pour valider les résultats de l'autre, ou de manière plus poussée en cherchant à obtenir une coopération plus fine. Nous avons donc repris l'ensemble de la chaîne de traitement de la plate-forme RECITAL pour en identifier les différentes phases et étudier pour chacune d'elles dans quelle mesure elle pourrait se prêter à une approche de coopération humain/algorithme.

Nous avons identifié trois niveaux d'intégration pour cette coopération. Le premier niveau donne à l'IA le rôle d'*assistant à l'annotateur*, qui propose à l'utilisateur (humain) des informations pertinentes afin de l'assister dans sa tâche - mais c'est bien l'humain qui prend les décisions et saisit les informations. Le second niveau consiste en une *assistance à l'expert* chargé de qualifier et valider les informations, offrant des signaux provenant tant des données saisies que des traces d'activité des utilisateurs sur la plate-forme visant à qualifier la qualité des informations qu'ils produisent. Ces signaux permettent à l'expert d'évaluer les informations, voire d'en proposer des corrections. Enfin, un troisième niveau envisagé consiste à intégrer l'IA sous forme d'*agent autonome* dans le système, en tant qu'annotateur ou qu'expert, qui participerait alors aux côtés des acteurs humains à la production et la validation collaborative des informations, interagissant en autonomie comme un utilisateur standard avec la plate-forme RECITAL pour y saisir des données.

Nous avons commencé à mettre en oeuvre les deux premiers niveaux au sein de la plate-forme RECITAL, exploitant les données déjà saisies. Le troisième niveau requiert des investigations plus poussées que nous poursuivons, tant sur ses principes que sa faisabilité technique dans son intégration à la plate-forme.

Références

Granet, Adeline, Benjamin Hervy, Geoffrey Roman-Jimenez, Marouane Hachicha, Emmanuel Morin, Harold Mouchère, Solen Quiniou, Guillaume Raschia, Françoise Rubellin et Christian Viard-Gaudin, 2018. Crowdsourcing-based Annotation of the Accounting Registers of the Italian Comedy. LREC.

Hervy, Benjamin, Pierre PÉTILLON, Hugo PIGEON et Guillaume RASCHIA, 2019. Correction des données : retour d'expérience sur la plate-forme RECITAL de transcription participative. In About Variety in Humanities Big Data, Recherche d'information, document et web sémantique. Vol. 19, No. 1, ISTE OpenScience.

Rubellin, Françoise et Guillaume Raschia, 2020. Redécouvrir les Théâtres de la Foire et la Comédie-Italienne avec les bases THEAVILLE et RECITAL. Revue d'Historiographie du Théâtre. Numéro 5. Trimestre 1.

" Site RECITAL ". 2021.Université de Nantes. 4 janvier. https://recital.univ-nantes.fr/

" Projet CIRESFI ". 2021. CETHEFI. 4 janvier. http://cethefi.org/ciresfi/