Extraction d'informations de bout en bout dans des documents manuscrits : vers une compréhension des actes de mariage de Paris de 1880 à 1940

L’extraction d’informations à partir de documents manuscrits repose traditionnellement sur trois étapes distinctes : l’analyse de la mise en page, la reconnaissance de texte manuscrit et l’extraction des informations clés telles que les entités nommées. Les approches récentes tentent de simplifier ce processus en adoptant des architectures dites end-to-end, intégrant ces étapes en une seule. Cependant, ces solutions unifiées peinent encore à atteindre les performances des modèles de langue appliqués aux textes numériques.

Dans cette présentation, nous introduisons DANIEL (Document Attention Network for Information Extraction and Labelling), une architecture end-to-end innovante, conçue pour une compréhension globale des documents manuscrits. DANIEL intègre un modèle de langue et effectue simultanément l’analyse de la mise en page, la reconnaissance de texte manuscrit et l’extraction d’entités nommées, et ce, directement sur des documents pleine page. Par ailleurs, cette architecture est capable d’apprendre de manière multi-tâche et multi-lingue, tout en s’adaptant à des mises en page variées.

Dans le cadre du projet EXO-POPP, notre ambition est de constituer une base de données complète de 300 000 actes de mariage de Paris et de sa banlieue, datant de 1880 à 1940. Ces documents, répartis sur plus de 130 000 scans de doubles pages, contiennent jusqu’à 118 types d’informations distinctes à extraire à partir du texte manuscrit.

Lors de cette intervention, nous présenterons les défis et les solutions apportées dans la mise en œuvre du projet EXO-POPP, tout en illustrant les performances de l’architecture DANIEL sur ce corpus.

Thomas Constum

Il a obtenu en novembre 2024 un doctorat dédié à l'extraction d'informations dans des documents historiques grâce à l’utilisation de grands modèles multimodaux.

Il a notamment participé aux projets POPP et EXO-POPP. Le projet POPP visait à constituer une base de données issue des recensements de Paris durant l'entre-deux-guerres, tandis que le projet EXO-POPP s’est focalisé sur l’extraction d’informations à partir d’actes de mariage du département de la Seine (1880-1940).

Thomas Constum est actuellement consultant en intelligence artificielle appliquée aux documents historiques pour l’entreprise Historical Consulting.

25 mars 2025