Appel à Projets de Recherche Doctoraux (PRD)
Programme Doctoral SCAI 2020

Formulaire de candidature

Proposition de projet

Intitulé du Projet de Recherche Doctoral

Analyse de l'espace littéraire : apprentissage automatique et évaluation des systèmes de reconnaissance des entités nommées

Directeur de thèse porteur du projet

Nom	ROE
Prénom	Glenn
Titre	Professeur de littérature française et humanités numériques
Ecole Doctorale de rattachement	ED 019 Littératures françaises et comparée
Unité de recherche – Intitulé – code – Tutelles - Nom du directeur/trice d'unité	Centre d'étude de la langue et des littératures françaises (CELLF) UMR 8599, dir. Claude Rétat
Equipe de recherche au sein de l'unité – Intitulé – Nom du Responsable d'équipe	LabEx OBVIL, dir. D. Alexandre
Adresse professionnelle	28, rue Serpente, 75006 Paris
Email	glenn.roe@sorbonne-universite.fr
Téléphone	06 08 97 97 65
Doctorants actuellement encadrés par le directeur de thèse (préciser le nombre de doctorants et leur année de 1ere inscription)	1. Ines Basille (2018) 2. Angélique Allaire (2019) 3. Axel Le Roy (2019) 4. Jean-Baptiste Tanguy (2019) 5. Hazar Massassati (2019)

Co-encadrant

Nom	LEJEUNE
Prénom	Gaël
Titre	Maître de conférences en informatique
HDR	Non (en cours d'écriture, soutenance prévue en octobre 2020)
Ecole Doctorale de rattachement	ED 433 Concepts et Langages
Unité de recherche – Intitulé – code – Tutelles - Nom du directeur/trice d'unité	STIH (Sens Text Informatique Histoire), Resp: Jacques Dürrenmatt
Equipe de recherche au sein de l'unité – Intitulé – Nom du Responsable d'équipe	Linguistique Computationnelle
Adresse professionnelle	28, rue Serpente, 75006 Paris
Email	gael.lejeune@sorbonne-universite.fr
Téléphone	06 62 81 17 58
Doctorants actuellement encadrés par le directeur de thèse (préciser le nombre de doctorants et leur année de 1ere inscription)	1 : Jean Baptiste Tanguy (depuis le 01/11/2019)

Co-encadrant

Nom	ALRAHABI
Prénom	Motasem
Titre	Ingénieur de recherche, doctorat
HDR	Non
Ecole Doctorale de rattachement	ED 019 Littératures françaises et comparée
Unité de recherche – Intitulé – code – Tutelles - Nom du directeur/trice d'unité	Centre d'étude de la langue et des littératures françaises (CELLF) UMR 8599, dir. Claude Rétat
Equipe de recherche au sein de l'unité – Intitulé – Nom du Responsable d'équipe	LabEx OBVIL, dir. D. Alexandre
Adresse professionnelle	28, rue Serpente, 75006 Paris
Email	motasem.alrahabi@paris-sorbonne.fr
Téléphone	06 58 69 81 92
Doctorants actuellement encadrés par le directeur de thèse (préciser le nombre de doctorants et leur année de 1ere inscription)	

Résumé du projet de recherche doctoral

(2000 caractères, **en français ou en anglais**, susceptible d'être mis en ligne)

Titre / Title:

Analyse de l'espace littéraire : apprentissage automatique et évaluation des systèmes de reconnaissance des entités nommées

Résumé / Abstract:

Les techniques de traitement automatique des langues (TAL) appliquées à l'analyse spatiale des corpus littéraires sont porteuses de promesses pour l'analyse des corpus littéraires. Ce thème de recherche émergent vise notamment l'analyse des lieux dans les œuvres, leur représentation cartographique ainsi que leur rapport aux personnages, à travers une diversité d'auteurs, d'époques ou de courants littéraires.

Malgré l'apparition d'outils de TAL opérationnels, notamment grâce à l'apprentissage profond, la tâche principale de l'analyse spatiale, à savoir la reconnaissance des entités nommées (REN), reste un problème épineux pour la langue française. Une limite importante à la performance de ces outils se situe dans la variabilité des données, de sorte que les résultats sur d'autres langues que l'anglais restent décevants. Ce manque de robustesse à la variation est particulièrement criant dès lors qu'il est question des corpus littéraires (variations diachroniques, diatopiques, etc.).

À l'intersection du TAL, de l'IA et des humanités numériques (HN), ce projet s'intéresse d'abord à l'évaluation de différentes approches et outils existants pour la REN spatiales et leur applicabilité sur les données littéraires. Ce travail s'appuiera donc sur des outils existants, mais nécessitera aussi le développement d'outils propres et de données manuellement annotées. À partir d'un corpus de 3000 romans du XIX^{e} et XX^{e} siècles, nous poserons la question de la granularité des EN spatiales (rues, villes, régions), de leur nature (réels, imaginaires, disparus) et de leur désambiguïsation.

Ce projet sera porté par une direction interdisciplinaire entre chercheurs en HN, en TAL et en IA avec, d'une part, la question de l'évaluation et de la valeur ajoutée apportée aux utilisateurs finaux des outils, et d'autre part, des questions épistémologiques sur les difficultés rencontrées par les systèmes d'apprentissage à gérer la variabilité et l'inconnu.

Le profil recherché est un(e) étudiant(e) de niveau M2, de formation en apprentissage automatique ou en TAL, avec un intérêt pour les HN littéraires.

AVIS et VALIDATION de l'ECOLE DOCTORALE

Joindre en annexe un descriptif du projet de recherche doctoral avec références au format pdf (« NOM_IA_2020 » / 2 pages maximum)

Ce formulaire et le descriptif de deux pages du projet doivent être envoyés simultanément par e-mail à l'ED de rattachement et à progdoc-scai@listes.upmc.fr avant le 22 mars 2020.

Titre : Analyse de l'espace littéraire : apprentissage automatique et évaluation des systèmes de reconnaissance des entités nommées

Descriptif du projet (2 pages)

Les techniques de traitement automatique des langues (TAL) pour l'analyse spatiale des corpus littéraires ont connu ces dernières années un progrès considérable. Ce thème de recherche émergent vise généralement l'analyse des lieux dans les œuvres littéraires, de leur représentation spatiale ou cartographique ainsi que leur rapport aux personnages (déplacements, rencontres, émotions…), parfois à travers une diversité de langues, d'auteurs, d'époques ou de courants littéraires (Roe et Burrows, 2020 ; Alrahabi et al., 2020).

Cependant, et malgré l'apparition d'outils de TAL opérationnels, la tâche principale pour l'analyse spatiale, à savoir la reconnaissance des entités nommées (REN), reste un défi majeur pour les systèmes d'Intelligence Artificielle (IA), d'extraction d'information et de veille épidémiologique notamment, comme la crise actuelle sur le coronavirus l'a rappelé (Kamel et Boulos 2020). En effet, la performance des outils de REN (Manning et al. 2014, Honnibal et Montani 2020) est, malgré des progrès importants, en train de rencontrer un certain nombre de limites. D'une part, on observe un plafond de verre (Stanislawek et al. 2019) dans la mesure où, même en combinant les meilleurs résultats de tous les systèmes, une part non négligeable des entités reste indétectable. D'autre part, les résultats sont obtenus dans des conditions très particulières en termes de langue et de type de textes. De sorte qu'il existe encore à l'heure actuelle une certaine carence sur les travaux en REN dans d'autres langues que l'anglais (Lejeune et al. 2015, Rashimi et al. 2019), que l'adaptation au domaine reste un problème ouvert (Jia et al. 2019) tandis que la variation diachronique fait assez peu partie des préoccupations de la communauté TAL (Kew et al. 2019).

Cette difficulté à gérer la variété rend les approches état-de-l'art peu satisfaisantes pour appréhender les données traitées en Humanités Numériques (HN), avec notamment des variations diachroniques, diatopiques ou encore discursives. Ce fait a figuré tout récemment à l'ordre du jour des préoccupations de l'Action Prospective Humanités Numériques Spatialisées du GDR MAGIS[1], qui y a ajouté également la question de la granularité des entités extraites. Cette préoccupation rejoint celle de la communauté internationale exprimée par le *Special Interest Group on Spatial Information* et de sa branche traitant des *Geospatial Humanities*.[2]

À l'intersection du TAL, de l'IA et des humanités numériques (HN), ce projet de thèse s'intéressera tout d'abord à l'évaluation de différentes approches et outils existants pour la REN spatiales et leur applicabilité sur les données littéraires. Le travail de recherche s'appuiera donc sur des outils existants, mais nécessitera aussi le développement d'outils propres, de données manuellement annotées et un guide d'annotation d'EN pour les textes littéraires. La création d'un tel guide sera facilitée par une première ébauche réalisée dans le cadre d'une collaboration entre

[1] https://projet.liris.cnrs.fr/aphns-magis/
[2] https://bgmartins.github.io/sigspatial-geohumanities/

l'OBVIL, l'EHESS et l'université Paul Valéry (Montpellier)[3] visant l'évaluation des guides existants dans le domaine de REN et la possibilité de leur utilisation sur des textes littéraires. Enfin, nous nous poserons la question de la granularité des EN spatiales (rues, villes, régions), de leur nature (lieux réels, lieux imaginaires, lieux disparus) et de leur désambiguïsation (*Named Entity Linking*).

Le corpus de travail sera composé d'un grand nombre de romans et récits de voyage du XIX^e^ et XX^e^ siècles (3 000 documents issus du projet ANR Chapitres, Paris 3 Sorbonne Nouvelle), et du corpus de la Très Grande Bibliothèque (35 000 documents issus d'un projet entre l'OBVIL et la BNF[4]). Ce dernier fait d'ores et déjà l'objet d'une thèse en cours dirigée par les deux laboratoires porteurs de cette proposition[5]. Elle pose la question de l'applicabilité des outils de TAL dans les contextes de corpus présentant des variations, et vise à améliorer la qualité de l'océrisation de notre corpus déjà collecté, dans l'objectif d'augmenter la performance des outils placés en aval de la chaîne de traitement (Abiven et al. 2020).

Les objectifs de la thèse seront, entre autres, de mettre à disposition de la communauté scientifique un modèle adaptatif pour le repérage des entités nommées spatiales dans les textes littéraires, un corpus standard annoté et un guide d'annotation pour les HN littéraires. Au-delà des HN, ce projet contribuera au travail de la communauté du TAL et de l'IA sur un certain nombre de verrous scientifiques actuellement rencontrés dans les données textuelles, et en particulier celui de la variabilité.

Ce projet nécessite une véritable collaboration interdisciplinaire entre chercheurs en HN, IA et TAL pour poser d'une part la question de l'évaluation et de la valeur ajoutée apportée par les outils aux utilisateurs finaux. D'autre part, cela soulève un certain nombre de questions épistémologiques sur les difficultés rencontrées par les systèmes d'apprentissage automatique dans la gestion de la variabilité et de l'inconnu.

La direction de cette thèse sera assurée par Glenn Roe, PR en littérature française et humanités numériques. Elle sera co-encadrée par Gaël Lejeune, MCF en informatique à SU, qui travaille particulièrement sur les questions de l'impact de la variation du matériau sur le TAL (en particulier le multilinguisme et bruitage, voir par ex. Baledent et al. 2020) et Motasem Alrahabi, ingénieur de recherche au LabEx OBVIL qui s'intéresse à l'application des modèles de TAL aux HN. Ce sujet associe différentes entités de la Faculté des Lettres : les laboratoires CELLF et STIH ainsi que le LabEx OBVIL. Le projet est donc adossé à une équipe de travail dynamique et pluridisciplinaire qui constituera un contexte fertile pour le doctorant recruté.

Le profil recherché est un(e) étudiant(e) de niveau M2, de formation en apprentissage automatique ou en TAL, avec un intérêt pour les HN littéraires.

---

[3] https://frama.link/GuideAnnotation

[4] http://api.bnf.fr/documents-de-gallica-produits-au-format-tei-par-obvil

[5] Contrat PhD[2] de la Région Île-de-France, "L'accessibilité et l'exploitation des documents textuels numérisés : un enjeu pour les bibliothèques numériques d'Île-de-France", Jean-Baptiste Tanguy.

Références :

K. Abiven, J.-B. Tanguy and G. Lejeune (2020). Acquisition semi-automatique et enrichissement de données en français pré-classique : variétés du français dans les mazarinades. Douzième congrès de l'Association des francoromanistes allemands (à paraître).

M. Alrahabi, C. Brando, M. Alkhalil, J. Dichy, (2020). Identifying and Analyzing Places in Arabic Travelogue Literature (à paraître).

A. Baledent, N. Hiebel and G. Lejeune (2020). Dating Ancient texts: an Approach for Noisy French Documents. Language Technologies for Historical and Ancient Langues (LT4HLA) (à paraître).

M. Boulos and E. Geraghty (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. International Journal of Health Geographics 19, Article number: 8.

M. Honnibal and I. Montani (2020). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (à paraître).

C. Jia, X. Liang and Y. Zhang (2019). Cross-Domain NER using Cross-Domain Language Modeling. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

T. Kew, A. Shaitarova, I. Meraner, J. Goldzycher, S. Clematide and M. Volk (2019). Geotagging a Diachronic Corpus of Alpine Texts: Comparing Distinct Approaches to Toponym Recognition. Proceedings of the Workshop on Language Technology for Digital Historical Archives.

G. Lejeune, R. Brixtel, A. Doucet and N. Lucas (2015). Multilingual Event Extraction for Early Epidemic Detection. Artificial Intelligence in Medicine, pp. 131-143.

C. Manning, M. Surdeanu, J. Bauer, J. Finkel, Prismatic Inc, S. Bethard, and D. Mcclosky (2014). The Stanford CoreNLP natural language processing toolkit. ACL demonstrations, 2014

A. Rahimi, Y. Li and T. Cohn (2019). Massively Multilingual Transfer for NER. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

G. Roe and S. Burrows, eds. (2020). Digitizing Enlightenment: digital humanities and the transformation of eighteenth-century studies. Oxford University Studies in the Enlightenment.

T. Stanislawek, A. Wróblewska, A. Wójcicka, D. Ziembicki and P. Biecek (2019). Named Entity Recognition - Is There a Glass Ceiling? Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL).