Offre de thèse : Structuration automatique des données de la littérature francophone

Période envisagée : sept 2020 - août 2023 (3 ans)

Description du sujet de thèse :

Le projet LIFRANUM vise à constituer et analyser le corpus des productions littéraires francophones nativement numériques. Il s'agit d'un projet financé par l'ANR qui regroupe un laboratoire de sciences humaines (MARGE), un laboratoire d'informatique (ERIC) et la Bibliothèque Nationale de France (BnF). Dans le cadre de ce projet, le laboratoire ERIC recherche un candidat pour une thèse de Doctorat qui débuterait en septembre 2020 pour une durée de 3 ans.

Constituer un corpus de productions littéraires nécessite d'être en mesure de structurer ce corpus de manière automatique afin d'aider l'utilisateur à accéder au contenu en fonction de ses besoins. Or ce corpus est un objet complexe qu'on peut voir comme un réseau d'information : il est composé de documents au contenu textuel parfois assez long (prose, poésie...) mais également de métadonnées (auteur, date...) et de liens connectant ces documents entre eux (par ex. via un hyperlien) voire les auteurs de ces documents (par ex. via une collaboration). L'une des tâches prise en charge par le laboratoire ERIC est précisément de proposer des nouvelles méthodes pour structurer un tel réseau d'information, en particulier par l'intermédiaire d'espaces de représentation adaptés. Les membres de l'équipe DMD ont récemment fait plusieurs contributions sur l'apprentissage de représentation de réseaux de documents [1,2,5,6].

L'objectif de la thèse qui est proposée consiste à construire des modèles innovants d'apprentissage de représentation et d'analyse de ce réseau d'information adapté au cas des données littéraires. Ces modèles doivent en particulier permettre de parcourir le corpus de manière originale, par exemple en capturant le "style" d'un auteur ou d'un groupe d'auteurs.

Plusieurs défis se posent alors :

- Le style d'un auteur ou d'un groupe d'auteurs est un concept mal défini et souvent difficile à identifier. Il ne se résume pas à un champ lexical ou à des considérations syntaxiques. Les modèles thématiques [9] ou les modèles distributionnels type word / sentence embeddings [8] ne suffisent pas pour représenter un auteur. Nous pensons qu'une combinaison de plusieurs indicateurs, l'auteur étant représenté comme un "mélange" sur différentes catégorisations, est une piste intéressante. Une autre piste est d'adopter une approche semi-supervisée incluant des contraintes capturée à partir des interactions avec les spécialistes en littérature.
- Les données relatives au projet ne sont pas disponibles en grande quantité. La structuration du corpus par des techniques de clustering peut tirer grandement

partie d'informations sémantiques au sujet des mots et des phrases apprises sur des grands corpus, par exemple à partir des modèles comme BERT [4] ou GPT-2 [10]. Le transfert des connaissances depuis ces modèles généralistes, en particulier ceux développés pour le français comme CamemBERT [7], risque de poser un certain nombre de problème lorsqu'il s'agit de le faire pour traiter de données issues de la littérature francophone. L'apport de connaissances issues de ressources comme des lexiques ou dictionnaires est une première solution à ce problème [11], mais leur combinaison avec les représentations contextuelles est loin d'être triviale.

- Pour finir, les partenaires du projet en SHS souhaitent pouvoir comprendre les raisons du rapprochement de tel auteur avec tel autre auteur, ou pourquoi un groupe d'auteurs (*cluster*) a été constitué. Chercher une "bonne explication" est une tâche difficile au coeur du développement d'IA "explicables" (XAI pour *eXplainable Artificial Intelligence*). Intégrer des connaissances issues de modèles distributionnels risque de la rendre encore plus difficile. Des membres de l'équipe ont déjà réalisé des contributions sur ce type de sujet mais pas nécessairement dédié au cas des données textuelles [3].

Le/la doctorant.e sera pleinement intégré.e au projet LIFRANUM et, à ce titre, devra participer à certaine des réunions de travail avec les partenaires. A ce titre, le candidat devra justifier d'un niveau satisfaisant de compréhension de la langue française. Il/elle sera en particulier en contact étroit avec un ingénieur de recherche qui sera recruté au laboratoire ERIC sur le projet courant 2021.

Lieu de la thèse

La thèse se déroulera dans les locaux du laboratoire ERIC, sur le campus de Bron de l'Université Lyon 2 (cf.: https://eric.msh-lse.fr/presentation/acces/).

Profil recherché

Étudiant.e ayant validé un Master en Informatique ou Mathématiques appliquées avec des compétences solides en science des données / machine learning. Des connaissances en NLP seront un vrai plus.

Dossier à envoyer

- CV
- lettre de motivation
- derniers relevés de note
- lettre(s) de recommandation

Contact

Julien.Velcin@univ-lyon2.fr

Références

[1] R. Brochier, A. Guille, J. Velcin: Global Vectors for Node Representation, WWW 2019.

[2] R. Brochier, A. Guille, J. Velcin: Inductive Document Network Embedding with Topic-Word Attention, ECIR 2020.

[3] I. Davidson, A. Gourru, S. Ravi: The Cluster Description Problem - Complexity Results, Formulations and Approximations, NeurIPS 2018.

[4] J. Devlin, M.W. Chang, K. Lee, K. Toutanova: Bert: Pre-training of deep bidirectional transformers for language understanding, NAACL-HLT 2019.

[5] A. Gourru, J. Velcin, J. Jacques, A. Guille: Document Network Projection in Pretrained Word Embedding Space, ECIR 2020.

[6] A. Gourru, J. Velcin, J. Jacques: Gaussian Embedding of Linked Documents from a Pretrained Semantic Space, IJCAI 2020.

[7] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, E.V. de la Clergerie... B. Sagot: CamemBERT: a Tasty French Language Model. arXiv preprint arXiv:1911.03894, 2019.

[8] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean: Distributed representations of words and phrases and their compositionality, NeurIPS 2013.

[9] J. Velcin, A. Gourru, E. Giry-Fouquet, C. Gravier, M. Rocher, P. Poncelet: Readitopics : Make your topic models readable via labeling and browsing, IJCAI 2018.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever: Language Models are Unsupervised Multitask Learners, Preprint on ArXiv 2019.

[11] J. Tissier, C. Gravier, A. Habrard: Dict2vec : Learning Word Embeddings using Lexical Dictionaries, EMNLP 2017.