Offre n°2025-08767

Post-Doctorant F/H Chercheur postdoctoral en IA responsable pour le journalisme

Niveau de diplôme exigé : Thèse ou équivalent

Fonction : Post-Doctorant

A propos du centre ou de la direction fonctionnelle

Le centre de recherche Inria de Saclay a été créé en 2008. Sa dynamique s'inscrit dans le développement du plateau de Saclay, en partenariat étroit d'une part avec le pôle de l'**Université Paris-Saclay** et d'autre part avec le pôle de l'**Institut Polytechnique de Paris**. Afin de construire une politique de site ambitieuse, le centre Inria de Saclay a signé en 2021 des accords stratégiques avec ces deux partenaires territoriaux privilégiés.

Le centre compte , dont 27 sont communes avec l'Université Paris-Saclay ou l'Institut Polytechnique de Paris. Son action mobilise **plus de 600 personnes** , scientifiques et personnels d'appui à la recherche et à l'innovation, issues de 54 nationalités.

Contexte et atouts du poste

Chaque année, la Direction des Relations Internationales d'Inria propose quelques postes postdoctoraux afin de soutenir les collaborations internationales.
Le contrat postdoctoral aura une durée de 18 mois. La date de début est entre le 1er juillet et le 1er septembre, mais pas plus tard que le 1er septembre.

Équipe:
Un chercheur postdoctoral potentiel intégrerait l'équipe Inria CEDAR tout en visitant l'équipe Human-Centered Data Analytics du CWI à Amsterdam. Ce projet est une collaboration avec les PI suivants :

Oana Balalau est chercheuse Inria au sein de l'équipe CEDAR, au centre Inria de l'Institut Polytechnique de Paris. Ses intérêts de recherche portent sur le traitement du langage naturel, en particulier sur la fouille d'argumentation, l'extraction d'informations et le data2text. Elle collabore avec des journalistes de plusieurs agences de presse : Radio France, Le Monde et AEF Info.
Davide Ceolin est chercheur au CWI au sein du groupe Human-Centered Data Analytics. Ses recherches portent sur la prévision transparente de multiples aspects de la qualité de l'information. Il est membre du laboratoire IA, médias et démocratie, un laboratoire multidisciplinaire qui étudie en profondeur les effets et les implications de l'IA pour les médias et la démocratie. Le laboratoire rassemble des chercheurs en informatique, droit et communication, ainsi que plusieurs partenaires de la société civile et industriels.

Les candidats intéressés peuvent contacter Oana Balalau s'ils ont des questions (oana.balalau@inria.fr).

Mission confiée

Les candidats aux postes postdoctoraux sont recrutés après la fin de leur doctorat ou après un premier post-doctorat : pour les candidats ayant obtenu leur doctorat dans l'hémisphère Nord, la date de la soutenance de la thèse sera après le 1er septembre 2022 ; dans l'hémisphère Sud après le 1er avril 2022. Afin de favoriser la mobilité, le poste postdoctoral doit se dérouler dans un environnement scientifique véritablement différent de celui du doctorat (et, le cas échéant, du poste occupé depuis le doctorat) ; une attention particulière aux candidats français ou internationaux ayant obtenu leur doctorat à l'étranger.

Contexte : Des systèmes de recommandation aux grands modèles de langage, les outils d'IA ont montré différentes formes de limitations et de biais [BHA+21, MMS+21, NFG+20]. Les biais dans les outils d'IA peuvent provenir de plusieurs facteurs, notamment les biais dans les données d'entraînement des outils d'IA, les biais de l'algorithme et les personnes responsables de la conception des outils d'IA, et les biais dans l'évaluation et l'interprétation des résultats des outils d'IA [NFG+20]. Les limitations sont dues à des difficultés techniques dans la réalisation de tâches spécifiques [SB22]. Les médias utilisent différentes aides algorithmiques dans leur travail : extractions d'entités et de relations, extraction d'événements, analyse des sentiments, résumé automatique, production semi-automatique des nouvelles à l'aide de modèles de génération de texte, et la recherche guidée par l'IA, entre autres [TJM+ 22, UBM23]. Compte tenu de l'importance du secteur des médias pour nos démocraties, des problemes dans les outils qu'ils utilisent pourraient avoir de graves conséquences.

Principales activités

Sujet de recherche:

Quelles sont les sources potentielles de biais dans les applications de traitement du langage naturel (TAL) destinées au journalisme et comment pouvons-nous les mettre en évidence et atténuer leurs effets ?

Pour répondre à cette question, nous étudierons deux cas d'utilisation.

Biais et limites dans les tâches de classification. Nous avons développé une plateforme de vérification des faits grâce à laquelle les journalistes peuvent suivre les déclarations des hommes politiques sur les réseaux sociaux [BEG+22]. Les déclarations les plus susceptibles d'être vérifiables sont mises en évidence, et pour cela, nous avons utilisé un algorithme d'apprentissage automatique. Les affirmations vérifiables (en anglais checkworthy) sont définies comme des phrases factuelles dont le grand public voudra savoir si elles sont vraies [HAL+17]. Notons que cette définition s'appuie sur ce qu'un annotateur considère comme étant d'intérêt général. De plus, l'ensemble de données d'entraînement contient des déclarations politiques. Par conséquent, les annotateurs pourraient avoir introduit par inadvertance un biais politique dans leurs annotations, par exemple en qualifiant plus souvent des phrases dignes d'être vérifiées s'elles sont exprimées par une personne d'une affiliation politique différente de la leur. Un deuxième modèle utilisé dans notre pipeline est la détection de la propagande, où la propagande est définie comme un ensemble de techniques de communication conçues pour influencer un lecteur et non pour l'informer. Les arguments fallacieux, qui sont des arguments incorrects que les vérificateurs de faits devraient démystifier, sont particulièrement intéressants. Alors que les définitions de la propagande sont plus précises en fonction du type exact de technique (par exemple, langage chargé, ad hominem), les ensembles de données annotés ont souvent un faible accord entre annotateurs [DSB+19]. En outre, les ensembles de données ne contiennent également que des déclarations politiques – encore une fois, un annotateur pourrait être plus enclin à qualifier de propagande le discours d'une personne ayant une opinion politique différente. Nous aimerions déterminer si ces ensembles de données et ces modèles sont biaisés et, si tel est le cas, étudier comment il pourrait être possible de mettre en évidence ce biais. Une idée intéressante consiste à intégrer le désaccord dans une tâche de classification en fournissant une explication textuelle de la raison pour laquelle un certain paragraphe pourrait avoir deux ou plusieurs étiquettes différentes (également connue en ML sous le nom de classification multi-étiquettes) selon deux ou plusieurs opinions humaines différentes. Comme mentionné, le désaccord pourrait venir de la définition de la tâche mais aussi des convictions des annotateurs. Cela implique de repenser le processus d'annotation, la formation et l'évaluation d'un modèle TAL, ainsi que la manière dont un modèle est utilisé pour une application réelle. Nous notons que le problème de la variabilité et des biais dans l'annotation humaine retient de plus en plus l'attention dans la communauté TAL [P22, UFH+21].

Biais et limites dans les tâches génératives. De nos jours, les modèles linguistiques génératifs sont utilisés pour diverses tâches, notamment pour des essais ou des textes argumentatifs. Nous en avons discuté avec des journalistes, qui ont confirmé qu'ils utilisaient de tels outils pour accélérer leur travail. Nous souhaitons nous concentrer sur des textes argumentatifs, notamment sur des sujets controversés dans notre société. Pour étudier le biais potentiel des modèles argumentatifs lorsqu'on leur demande de fournir des informations sur de tels sujets, nous aimerions comparer les textes argumentatifs générés automatiquement avec des textes argumentatifs issus du crowdsourcing, tels que les textes hébergés sur les plateformes de débat. Ce projet peut être étendu à l'analyse de la manière dont les sujets controversés sont débattus dans la sphère publique, par exemple en se concentrant sur les débats des campagnes électorales en cours. Le premier défi technique de cette tâche consiste à identifier des arguments similaires - lorsqu'un argument est composé d'une affirmation et des preuves à l'appui de cette affirmation. La même affirmation peut être étayée par différentes preuves, et il est également important de mettre en évidence ces différences, car une préférence pour un certain type de preuves pourrait montrer des tendances plus importantes. Par exemple, l'affirmation « L'avortement devrait être légal ». peut être soutenu par « Une femme devrait toujours avoir le choix sur son corps ». ou la phrase « Dieu nous a donné le libre arbitre et nous devons respecter le libre arbitre des autres ». Un deuxième défi technique consiste à mesurer le degré de persuasion d'un texte argumentatif, par exemple en mesurant le degré d'exhaustivité des preuves présentées [HG16].

Les références:

[BEG+22] Balalau, O., Ebel, S., Galizzi, T., Manolescu, I., Massonnat, Q., Deiana, A., Gautreau, E., Krempf, A., Pontillon, T., Roux, G. and Yakin, J., 2022, October. Fact-checking Multidimensional Statistic Claims in French. In *TTO 2022-Truth and Trust Online*.

[BHA+21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

[DSB+19] Da San Martino, G., Seunghak, Y., Barrón-Cedeno, A., Petrov, R. and Nakov, P., 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5636-5646). Association for Computational Linguistics.

[HAL+17] Hassan, N., Arslan, F., Li, C. and Tremayne, M., 2017, August. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1803-1812).

[HG16] Habernal, I. and Gurevych, I., 2016, November. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1214-1223).

[MMS+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*,54(6):1–35, 2021.

[NFG+20] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356,2020.

[P22] Plank, B., 2022, December. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 10671-10682).

[SB22] Chirag Shah and Emily M Bender. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 221–232, 2022.

[TJM+22] Christoph Trattner, Dietmar Jannach, Enrico Motta, Irene Costera Meijer, Nicholas Diakopoulos, Mehdi Elahi, Andreas L Opdahl, Bjørnar Tessem, Nj?al Borch, Morten Fjeld, et al. Responsible media technology and ai: challenges and research directions. *AI and Ethics*, *2*(4):585–594, 2022.

[UBM23] Prajna Upadhyay, Oana Balalau, and Ioana Manolescu. Open information extraction with entity focused constraints. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1255–1266, 2023.

[UFH+21] Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Plank, B. and Poesio, M., 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, *72*, pp.1385-1470.

Compétences

Compétences techniques et niveau requis : solides connaissances en TAL et bonnes compétences en programmation

Langues : Anglais

Avantages

- Restauration subventionnée
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle

Rémunération

Selon expérience

Informations générales

- **Thème/Domaine :** Représentation et traitement des données et des connaissances
  Statistiques (Big data) (BAP E)
- **Ville :** Palaiseau
- **Centre Inria :** Centre Inria de Saclay
- **Date de prise de fonction souhaitée :** 2025-07-01
- **Durée de contrat :** 1 an, 6 mois
- **Date limite pour postuler :** 2025-06-30

Contacts

- **Équipe Inria :** CEDAR
- **Recruteur :**
  Balalau Oana-denisa / oana.balalau@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers

différents. 900 personnels d’appui à la recherche et à l’innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

Le candidat doit soumettre :

- CV détaillé avec une description du doctorat et une liste complète des publications avec les deux plus significatives mises en évidence
- Lettre de motivation
- 2 lettres de recommandations
- Copie du passeport

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.