INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

# Campagne « Contrats Doctoraux Politique Doctorale Etablissement » - INSA RENNES 2022

**FICHE DESCRIPTIVE (3 pages maximum*) – annexes autorisées (pdf)**

| | |
|---|---|
| **Catégorie demande** | ☒ contrat doctoral 100 %<br>☐ contrat doctoral 100 % situation handicap (compléter et joindre le dossier de candidature joint)<br>☐ contrat doctoral 50 % |
| **Si demande effectuée sur l'appel Contrats Doctoraux Ordinaires indiquez-le** | ☒ contrat doctoral 100 %<br>☐ contrat doctoral 50 % |
| **Si cofinancement nature du cofinancement**** | **(partenaire/organisme, montant cofinancement, nature du contrat, durée, calendrier)** |
| **Titre du sujet de thèse** | Détection de relations à partir de peu ou pas d'exemples dans les archives de presse |
| **Unité/équipe** | ☐ Inst. FOTON<br>☐ ISCR-CSM<br>☐ ISCR-CM<br>☐ ISCR-COrint<br>☐ IETR<br>☒ IRISA<br>☐ IRMAR<br>☐ LGCGM<br>☐ LS2N/MECAPROCE |
| **Direction de la thèse/Codirection (HDR)** | Pascale Sébillot / Guillaume Gravier |
| **Co-encadrement** | Michel Le Nouy (Ouest France) |
| Partenaires impliqués | Ouest-France |
| **Cotutelle / co-direction internationale envisagée* ?** | non |
| **Description des objectifs scientifiques du projet de thèse** | La thèse s'inscrit dans une collaboration entre l'IRISA et le journal Ouest-France et vise de manière générale à développer des approches en traitement automatique des langues (TAL) utilisables en pratique dans le contexte du journal.<br><br>Ouest-France dispose d'une base de contenus couvrant de nombreuses publications du groupe, dont les multiples éditions de Ouest-France, et regroupant actuellement plus de 35 millions de documents sur des sujets divers. Cette base de contenus contient des informations et connaissances précieuses dont une partie peut être extraite de manière automatique. En particulier, ces archives font état au sein des articles d'entités (personnes, lieux, organisations) et de nombreuses relations qui peuvent exister entre elles (est maire de, est situé à, a démissionné de, a son port d'attache à, etc.). Ces relations sont particulièrement importantes pour les journalistes, par exemple pour donner du contexte à de nouveaux articles, trouver des sujets nouveaux ou des angles de rédaction originaux. Il est donc important de pouvoir les détecter de manière automatique par des techniques de TAL permettant l'extraction de relations (détection des entités, détection de l'existence d'une relation, classification de la nature de la relation – *cf.* (Nasar et al., 2021)).<br><br>L'état de l'art en extraction de relations s'appuie cependant sur des approches gourmandes en données d'apprentissage, en particulier pour l'étape de classification. De plus, les modèles de classification sont appris pour des classes génériques (est situé à, est dirigeant de, etc.) présentes dans les jeux de données standards comme TACRED (Zhang et al., 2019), ce qui ne correspond que très partiellement aux besoins des journalistes. En effet, ces derniers peuvent s'intéresser à des relations rares (signaux faibles) et leurs centres d'intérêt évoluent constamment (a démissionné de, a son port d'attache à). L'enjeu aujourd'hui est donc de passer de la détection et classification de relations pour lesquelles on peut avoir des modèles appris sur de grands volumes de données, à des relations plus rares pour lesquelles on ne dispose que de quelques exemples, souvent un, voire d'aucun.<br><br>L'objectif de la thèse est d'étudier l'apport des techniques d'apprentissage frugales en données (*zero shot* et *few shot learning*) dans ce contexte (Wang et al., 2019 ; Wang et al., 2020). Ces techniques, en particulier le *zero shot learning*, requièrent la définition d'un prototype (un vecteur) permettant de faire le lien entre une description sémantique de ce que l'on souhaite reconnaître et le contenu dans lequel on souhaite reconnaître |

| | |
|---|---|
| | ce concept (Le Cacheux, 2020) : dans notre cas, le prototype est un vecteur (*embedding*) qui décrit une relation entre deux entités, que l'on peut relier aux contenus où les relations sont mentionnées et que l'on peut obtenir à partir d'un exemple (*few shot*) ou d'une description quelconque (*zero shot*). Dans ce dernier cas, les approches cherchent à comparer le prototype obtenu à partir de la description de la relation (ou d'un exemple) avec les prototypes de relations bien connues pour lesquelles nous disposons de modèles pré-entrainés, afin de déterminer à la volée un modèle de la nouvelle relation. Si ces approches correspondent bien à notre scénario d'utilisation, les techniques de *few shot* et *zero shot learning* sont souvent étudiées pour des domaines restreints où il est aisé de donner une description de l'objet que l'on veut détecter et de définir un prototype. Elles restent peu appliquées à des contextes plus difficiles, notamment à l'extraction de relations (*cf.* (Levy et al., 2017 ; Yuan et al., 2017 ; Chen et al., 2021) pour des points d'entrées sur le sujet).<br><br>Le travail s'articulera autour de trois grandes questions : Comment représenter une relation entre deux entités sous la forme d'un vecteur ? Comment décrire une relation non connue pour en donner un prototype ? Quel apport, quelle utilité pour la fouille d'archives par les journalistes ?<br><br>Sur le premier point, nous avons développé une première approche utilisant les RNN pour encoder le plus court chemin de dépendances (syntaxiques) entre les deux entités, que nous avons évaluée de manière qualitative sur une tâche de *clustering* de relations. On pourra partir de cette approche pour en mener dans un premier temps une évaluation quantitative dans différents contextes (*clustering*, classification, etc.) avant de chercher à la rendre plus performante (réorganisation syntaxique, meilleure abstraction, modèles d'*embedding* plus performant (BERT-*like*), etc.). Sur le deuxième point, l'idée est d'apprendre des modèles permettant de passer d'une description textuelle courte de la nature de la relation, telle que fournie par un journaliste, à son prototype afin de s'affranchir d'un exemple de la relation. En l'absence de données pour apprendre un tel modèle, on pourra exploiter les techniques récentes de génération de texte afin de générer des données synthétiques (texte décrivant une relation) pour des relations dont le prototype et la nature sont connus. Une autre piste consiste à tirer profit du contexte Ouest-France pour créer un corpus de description de relations par un journaliste afin d'apprendre le modèle permettant de passer de la description au prototype. Enfin, la thèse se déroulant dans le cadre d'un partenariat étroit avec Ouest-France, l'évaluation de l'apport des technologies développées se fera dans le cadre d'expérimentation *in situ* avec le service R&D de Ouest-France et les journalistes de la rédaction. |
| **Adéquation du projet avec le critère 1** | **Le groupe SIPA-Ouest-France est un groupe national majeur du domaine des médias.** Le journal *Ouest-France* est le quotidien national le plus lu en France (58 bureaux éditoriaux dans 12 départements ; 40 éditions quotidiennes spécifiques à des territoires, 10 éditions le dimanche). Le groupe diffuse aussi 4 autres quotidiens régionaux (dont *Presse-Océan*), ainsi que des hebdomadaires locaux et des magazines ; il est également co-éditeur de *20 minutes*.<br><br>**L'équipe Linkmedia de l'IRISA**, dont l'objet est le traitement de très larges collections de documents multimédias, **entretient depuis quelques années des relations étroites avec le groupe Ouest-France**. Une thèse CIFRE Insa (co-dirigée par P. Sébillot et G. Gravier, ainsi que M. Le Nouy côté Ouest-France) doit se soutenir au cours de l'année 2022, et le contrat Inria Project Lab iCODA (2017-2021), porté par Linkmedia, réunissait en particulier Ouest-France et 4 équipes Inria. Enfin Linkmedia est actuellement en phase de réflexion et de discussions pour monter avec ce partenaire un laboratoire commun de recherche (LCR CNRS) et un Labcom (ANR).<br><br>**La demande s'inscrit dans l'objectif d'un renforcement du partenariat que l'équipe Linkmedia entretient avec l'entreprise Ouest-France, majeure dans son domaine des médias**. Le sujet de la thèse a pour objet de faciliter l'accès à l'information contenue dans les archives du groupe, afin d'offrir des outils – développés en répondant à des questionnements scientifiques novateurs – qui soient effectivement utilisables par les journalistes. **Le sujet aura donc des retombées directes pour l'entreprise**.<br><br>**Le sujet de thèse s'inscrit pleinement dans le DIS Économie numérique sécurisé et responsable de la S3**, et en particulier dans ses objectifs stratégiques 1 (Booster l'innovation dans les technologies et applications liées au numérique) et 2 (Construire une intelligence collective autour de la donnée). Il concerne d'une part les leviers *Images et contenus / réseaux et IoT* et *Données et intelligence*, s'intéressant au stockage (sous |

| | |
|---|---|
| | forme d'archive de données numériques média, et de graphe des connaissances extraites), au traitement et à la valorisation de contenus (réutilisation pour la production de nouveaux contenus contextualisés grâce aux informations extraites) ; il s'inscrit aussi dans le levier *Technologies numériques sobres et responsables*, l'utilisation de techniques *zero-shot learning* visant à limiter les volumes de données nécessaires à l'apprentissage des modèles, se positionnant dans une optique d'IA frugale. **Le sujet s'inscrit dans une moindre mesure dans le DIS Économie de l'industrie pour une production intelligente,** dans son objectif stratégique 2 (Développer une industrie de transition) et son levier *L'humain dans l'industrie des usages*, son objectif étant de proposer aux journalistes des outils centrés sur leurs besoins réels, visant à leur faciliter l'écriture de nouveaux articles.<br><br>**Pour ce qui est de la SNR, le sujet se positionne naturellement au cœur du Défi 7 Société de l'information et de la communication** (orientation 28 : Exploitation des grandes masses de données) et du programme d'action Big Data. **Il s'ancre également dans le Défi 3 (Le renouveau industriel),** orientation 11 (Usine numérique), par son utilisation des outils numériques pour améliorer les fonctions et performances de l'industrie du multimédia. Ses aspects IA frugale entre en résonance avec le **Défi 1** (Gestion sobre des ressources et adaptation au changement climatique) par son aspect gestion sobre des ressources impactées par l'activité industrielle humaine. |
| **Adéquation du projet avec le critère 2** | **(Qualité de l'encadrement et de l'environnement de recherche)**<br>Le doctorant ou la doctorante sera accueilli au sein de l'équipe Linkmedia de l'IRISA, spécialiste du traitement et de la valorisation de très larges collections de documents multimédias à l'aide de techniques d'apprentissage automatique. Il ou elle sera aussi en contact étroit avec les partenaires Ouest-France de l'équipe. Les 2 co-directeurs de thèse proposés ont une longue expérience commune d'encadrements (dont 6 thèses soutenues, 2 en cours) et possède une bonne visibilité internationale sur le sujet dans les communautés TAL et multimédia. |
| **Adéquation du projet avec le critère 5** | **(Lien avec les objectifs de responsabilités sociétales et développement durable)**<br>ODD Nations Unies<br>Le sujet de thèse se positionne, de par sa thématique liée à l'industrie des médias, au sein de **l'Objectif 9** (Industrie, innovation et infrastructure). Son but est d'extraire des connaissances à partir de contenus journalistiques pour aider à la mise en perspective et en contexte de nouveaux contenus à produire. Il a donc pour objet d'accroître la capacité des journalistes à analyser, dans un contexte vaste, l'information, afin de fournir à tous des informations fiables et étayées répondant au droit à une information véridique et vérifiable, ce qui l'ancre, au-delà de ce premier objectif, également dans dans l'**Objectif 16** (Paix, justice et institutions efficaces), voire l'Objectif 10 (Inégalités réduites). Enfin les méthodes d'extraction de connaissances à partir de contenus reposent principalement sur des techniques d'apprentissage à partir de masses de données, annotées ou non. L'apprentissage des modèles à partir de ces données conséquentes est particulièrement gourmande en énergie. L'emploi de technique *zero-shot learning* a, entre autres, pour objectif de limiter ce besoin de données annotées massives et d'obtenir des modèles en se fondant sur des modèles préexistants, déjà disponibles, s'inscrivant en cela dans une optique d'IA frugale. Les problématiques du sujet entrent donc en résonance avec **l'Objectif 13** (Lutte contre les changements climatiques). |
| **Autres informations utiles (planning envisagé…)**<br>- **Pièces jointes autorisées** | **Bibliographie**<br>Nasar, Z., Jaffry, S.W. and Malik, M.K. Named entity recognition and relation extraction: State-of-the-art. ACM Computing Surveys, 54(1) :1-39, 2021.<br><br>Zhang Z., Han X., Liu Z., Jiang X., Sun M., and Liu Q. ERNIE: Enhanced language representation with informative entities. Proc. Annual Meeting of the Association for Computational Linguistics, 2019.<br><br>Wang, W., Zheng, V.W., Yu, H. and Miao, C. A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology, 10(2):1-37, 2019.<br><br>Wang, Y., Yao, Q., Kwok, J.T. and Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys, 53(3):1-34, 2020.<br><br>Le Cacheux, Y. Vers un apprentissage sans exemple plus réaliste. Thèse du Conservatoire national des arts et métiers, 2020. |

| | |
|---|---|
| | Levy, O., Seo, M., Choi, E. and Zettlemoyer, L. Zero-shot relation extraction via reading comprehension. Proc. Conf. on Computational Natural Language Learning, 2017.<br><br>Yuan, J., Guo, H., Jin, Z., Jin, H., Zhang, X. and Luo, J. One-shot learning for fine-grained relation extraction via convolutional siamese neural network. Proc. IEEE International Conference on Big Data, 2017.<br><br>Chen, C.Y. and Li, C.T. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. |

**** joindre justificatifs associés***

Je déclare exactes toutes les informations contenues dans ce document
Date et signature du responsable scientifique du projet :
Le 25/01/2022



| Avis du directeur d'unité : | Avis du responsable de composante : |
|---|---|
| Date et signature : | Date et signature : |

Classement du responsable de composante