Offre d'emploi post-doctoral :
Extraction de connaissances à partir de textes et de tableaux pour enrichir un graphe de connaissances

Le projet ANR ECLADATTA (dont les partenaires sont l'IRIT, ORANGE-labs et EURECOM) démarre le 15 mars 2023. Ce projet s'inscrit dans la volonté d'enrichir de façon qualitative les graphes de connaissances comme Wikidata, DBpedia, etc. En effet, les graphes de connaissances représentent un enjeu majeur pour bon nombre d'applications comme la recherche de jeux de données, les systèmes de recommandation, les systèmes question/réponse, et plus généralement les systèmes basés sur les connaissances. Une des manières d'enrichir ces graphes est d'extraire des informations de textes, en particulier de pages web. Par exemple, une des sources majeures pour compléter un graphe de connaissances est de reprendre des graphes existants comme DBpedia, ou d'analyser les pages de Wikipedia ou plus largement de WIkimedia. Or ces pages contiennent à la fois du texte en langage naturel et des éléments plus structurés comme des tableaux

Dans ce contexte, **le projet ECLADATTA a pour objectif de combiner et de confronter les connaissances extraites du texte brut, des tableaux présents dans le texte et des graphes de connaissances, pour enrichir/corriger les graphes de connaissances**. Le rôle de l'IRIT dans ce projet est d'étudier la complémentarité des approches et des données extraites des textes d'une part, des tableaux d'autre part.

Par exemple, lorsqu'on tente d'extraire du texte, des entités et les liens sémantiques qui les relient, les données présentes dans les tables peuvent aider à cibler les entités pertinentes, leurs types et leurs propriétés. Les tableaux peuvent aussi aider à désambiguïser la nature d'une relation implicite entre deux entités mentionnées dans le texte, et ce grâce aux en-têtes des colonnes. À l'inverse, le texte autour du tableau peut fournir du contexte (e.g. comment, quand, où, pourquoi les données ont été produites), information absente des tableaux.

Par ailleurs, les données extraites de ces différentes sources peuvent être inconsistantes, voire contradictoires, comme par exemple lorsque la validité d'une connaissance dépend de conditions temporelles (voir les qualifiers de Wikidata) ou que la connaissance se trouve via des représentations différentes (voir les fonctions d'agrégation comme la somme ou la moyenne).

Dans le cadre du post-doc proposé, un corpus annoté sera fourni par les partenaires du projet, ainsi que des outils d'extraction déjà développés au sein de ces équipes et adaptés à ce corpus dans le cadre d'ECLADTA : extraction de désambiguïsation d'entités nommées (NERD et ADEL), extraction de relations à partir de texte (BizRel) et de tableau (DAGOBAH), et classification (ZeSTE). Ces outils sont basés sur des algorithmes d'apprentissage automatique. Les missions de la personne recrutée consisteront à :

- réaliser un état de l'art sur les travaux ayant abordé ces problématiques (extraction conjointe, confrontation des connaissances extraites, etc.)
- s'approprier le corpus, expérimenter les outils existants sur ce corpus de textes
- proposer des méthodes (nouvelles ou reprises de l'état de l'art) permettant de combiner les stratégies des outils existants
- proposer et implémenter des méthodes pour évaluer la cohérence et la complémentarité des connaissances extraites
- implémenter et évaluer de nouveaux outils intelligents basés sur ces méthodes

Compétences scientifiques attendues : Extraction de relations à partir de texte – TAL - Apprentissage automatique (modèles neuronaux)
Autres compétences requises : Maîtrise du français et de l'anglais - Autonomie et prise d'initiative - Capacité de rédaction et de communication - Rigueur et méthodologie

Environnement de travail
Le post-doctorat se déroulera au sein du laboratoire IRIT, à Toulouse, dans l'équipe MELODI qui comporte environ 40 personnes (réparties sur les sites des Universités Paul Sabatier et Jean Jaurès). MELODI est composée d'un groupe TAL et d'un groupe Ontologies et Web Sémantique.
Le télétravail est autorisé, néanmoins 2 jours sur place au minimum sont requis.
Le salaire est compris entre 26000€ et 36000€ annuel net selon l'expérience. Ce salaire comprend Sécurité Sociale et Cotisation Retraite. Le **poste est à pourvoir dès le 1 mai 2023**. Toutefois, le processus de recrutement prend environ 2 mois.

Contacts : Nathalie Aussenac-Gilles (aussenac@irit.fr) - Mouna Kamel (kamel@irit.fr) - Véronique Moriceau (veronique.moriceau@irit.fr)
Equipe MELODI – IRIT – Université Paul Sabatier – 118, route de Narbonne – 31062 Toulouse Cedex 9