POSTDOC
LLL-CNRS, Université d'Orléans, France

Titre: Extraction d'information : résolution de coréférences et extraction de relations temporelles

Date limite de candidature : 24 mars 2023
Début souhaité : à partir de mai 2023
Durée : 12 mois
Financement : Projet régional DOING (APR-IA, Region Centre Val de Loire)
Laboratoire de recherche : LLL-CNRS (https://lll.cnrs.fr/), Université d'Orléans, France
Collaboration avec les laboratoires suivants :
---- LIFAT (https://lifat.univ-tours.fr/) et
---- LIFO (https://www.univ-orleans.fr/lifo/?lang=en)
Salaire : 2580€ brut mensuel

CONTEXTE
Le projet régional DOING vise à développer des méthodes et des outils pour, dans un premier temps, extraire des informations de données textuelles en les structurant dans une base de données graphe, puis pour manipuler de façon intelligente ce graphe de connaissance. Le domaine d'application choisi est le domaine de la santé, avec en premier lieu l'utilisation de données disponibles librement (tels que des cas cliniques). DOING vise à concevoir une nouvelle forme de requêtes déclaratives, pouvant intégrer des analyses, qui guideront les spécialistes du domaine de la santé dans leur prise de décision. DOING est conçu sur une réelle collaboration interdisciplinaire (Traitement Automatique des Langues, Bases de Données et Intelligence Artificielle) pour transformer des données en information puis en connaissance.

DOING (https://www.univ-orleans.fr/lifo/evenements/doing/) est basé sur une collaboration inter-disciplinaire pour transformer des données en information puis en connaissance.
Ce projet s'inscrit dans le cadre de l'action DOING@MADICS et du groupe de travail DOING@DIAMS.

SUJET
Le travail de recherche concerne la tâche 1 du projet DOING, et se concentrera principalement sur le 2ème volet concernant l'extraction des relations.

(1) Détection et catégorisation des entités d'intérêt (e.g, pathologie, traitement). Un premier travail a été effectué pour développer un système à base de CRF (Minard et al., 2020) dans le cadre de notre participation à la campagne d'évaluation DEFT 2020. Une tâche consistera à améliorer ce système, éventuellement en utilisant d'autres corpus de cas cliniques : corpus de DEFT 2019, 2020 annotés avec des entités cliniques ; corpus E3C en cours d'annotation (Magnini et al., 2020).
(2) Détection et catégorisation des relations entre les entités. L'extraction des relations permet de faire émerger un sens qui sera représenté dans graphe de données que nous cherchons à construire. Nous nous focaliserons sur deux types de relations essentielles :
(2.1) Les relations de coréférences : elles existent entre deux unités linguistiques faisant référence à une même entité du discours (par exemple "Alice ressent des frissons. Ce symptôme est lié à sa fièvre."). Les techniques actuelles reposent sur des modèles neuronaux classiques, génériques, entraînés des corpus du domaine général, et souffrent d'une importante dégradation de performance lorsqu'ils sont appliqués à un autre domaine (Zhang et al., 2020). Tout en considérant la spécificité de la coréférence en langue de spécialité médicale, nous chercherons à développer des modèles relativement génériques, plus robustes en termes de performance inter-domaine et explicables. Pour lever ce dernier verrou, nous utiliserons des techniques d'apprentissage statistique interprétable (arbres de décision, forêts d'arbres aléatoires) proposées par notre groupe TAL (Desoyer et al.,

2014). En parallèle, nous pourrons adapter le système neuronal proposé par le laboratoire LATTICE dont le concepteur est associé au LIFO (Grobol, 2020).
(2.2) Les relations temporelles : elles permettent d'ordonner les événements concernant un patient (l'apparition des symptômes, les traitements suivis, etc.). Nous travaillerons au développement d'un système d'extraction d'informations temporelles et testerons des méthodes permettant de pallier la faible quantité de données dans le domaine médical. Nous poursuivrons également les travaux entamés dans les projets Temporal et ODIL par notre groupe TAL (Lefeuvre-Halftermeyer et al., 2016), en évaluant la généricité du schéma proposé et en étudiant l'ajout possible de la notion de containers (Pustejovsky and Stubbs, 2011).
(3) Interaction entre l'extraction des informations, la construction et l'exploitation des bases de données graphe. Une méthode viserait à considérer comment l'extraction des relations peut bénéficier des approches utilisées sur les bases de données graphe (page rank, betweenness, etc). Les observations issues de T3 viendraient compléter et s'intégrer aux résultats de T1 pour l'instanciation de la base.

ÉQUIPE
Ce postdoc s'inscrit dans le cadre d'un travail collaboratif qui implique les chercheurs suivants :
- LLL : Anne-Lyse Minard, Lotfi Abouda, Flora Badin, Emmanuel Schang
- LIFO : Anaïs Lefeuvre-Haltermeyer
- LIFAT : Jean-Yves Antoine
- LISN : Agata Savary

PROFIL RECHERCHÉ
Le candidat doit :
- posséder un doctorat en informatique ou en linguistique avec une spécialisation en TAL
- avoir des connaissances préalables en apprentissage automatique
- une expérience passée sur les thématiques suivantes sera appréciée : résolution de coréférences, extraction de relations temporelles, extraction d'information en domaine de spécialité
- avoir un niveau de français permettant d'analyser les données et d'échanger avec l'équipe

Le travail de recherche est mené au Laboratoire Ligérien de Linguistique (LLL) à Orléans. La personne recrutée devra être présente physiquement (il n'est pas possible de travailler à distance).

POUR POSTULER
Vous devez envoyer votre candidature par mail, au plus tard le 24 mars 2023, à l'adresse suivante : anne-lyse.minard@univ-orleans.fr
Le dossier de candidature doit contenir :
- un CV détaillé
- une lettre de motivation
- le diplôme de doctorat
- les rapports de pré-soutenance
- deux références

BIBLIOGRAPHIE
Grobol L., Coreference resolution for spoken French. Thèse Univ. Paris Sorbonne Nouvelle, 2020.
Lefeuvre-Halftermeyer A., Antoine J-Y., Couillault A., Schang E., Abouda L., Savary A., Maurel D., Eshkol I. and Battistelli D. Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility. In Proceedings of the 10th Int. Conf. on Language Resources and Evaluation (LREC), 2016.
Magnini B., Altuna B., Lavelli A., Speranza M., Zanoli R. The E3C Project:Collection and Annotation of a Multilingual Corpus of Clinical Cases. In Proceedings of CLiC-it 2020.

Minard A-L., Roques A., Hiot N., Halfeld-Ferrari M., Savary A. DOING@DEFT : cascade de CRF pour l'annotation d'entités cliniques imbriquées. Actes 27° conférence TALN, Nancy, France, 2020.
[16] Pustejovsky J. and Stubbs A. Increasing Informativeness in Temporal Annotation. In Proceedings of the 5th Linguistic Annotation Workshop. Ass. for Computational Linguistics, Stroudsburg, PA, USA, LAW V ’2011.
Zhang H., Zhao X., Song Y. A Brief Survey and Comparative Study of Recent Development of Pronoun Coreference Resolution – preprint arXiv:2009.12721, 2020.
Desoyer A., Landragin F., Tellier I., Lefeuvre A., Antoine J.-Y. Les coréférences à l’oral : une expérience d’apprentissage automatique sur le corpus ANCOR, Traitement Automatique des Langues, TAL, vol. 55 (2), 2014.