

Sujet de thèse

Titre de la thèse : Response generation models for solving multi-faceted information needs

Directrice ou directeur de thèse : Lynda Tamine (IRIT), Karen Pinel-Sauvagnat (IRIT), Laure Soulier (ISIR)

Collaboration dans le cadre de la thèse : CIFRE avec Ecovadis

Laboratoire d'accueil :

IRIT (Institut de Recherche en Informatique de Toulouse), Campus Paul Sabatier à Toulouse
Ou

ISIR (*Institut des Systèmes Intelligents et de Robotique*), Campus Pierre et Marie Curie, 4 place Jussieu, 75005 Paris.

Personne à contacter

Prénom Nom : Laure Soulier

Tel : +33 1 44 27 74 91

Email : laure.soulier@isir.upmc.fr ; tamine@irit.fr ; sauvagnat@irit.fr ; skatrenko@ecovadis.com

Envoyer votre candidature par mail, avec [CIFRE Ecovidis] en objet, un CV, une lettre de motivation, des relevés de notes du L3 au M2, ainsi que des lettres de recommandations

Date limite de dépôt de la candidature : 31/08/2022 (candidatures traitées au fil de l'eau, clôture quand nous trouvons un.e candidat.e

Description du sujet (en français)

Contexte :

La perspective de nouveaux systèmes de recherche d'information (RI) (par exemple, les systèmes conversationnels orientés recherche ou les systèmes prenant en charge des tâches de recherche complexes) a favorisé la recherche de modèles théoriques de recherche d'information qui tirent partie des interactions des utilisateurs ou les prennent en compte, par exemple, par la clarification des questions ou des modèles interactifs. Cependant, très peu de travaux se concentrent sur la manière d'interagir avec l'utilisateur en langage naturel, ce qui est essentiel, par exemple pour les systèmes conversationnels.

Description du projet :

L'objectif principal de la thèse est de concevoir des modèles de questions-réponses visant à résoudre des besoins d'information multi-facettes. Plus particulièrement, étant donné une collection de documents, notre objectif est de générer des réponses structurées et complètes, couvrant toutes les facettes d'un besoin d'information complexe.

Pour ce faire, des approches et des modèles issus de la recherche d'information (RI) et du traitement du langage naturel (TLN) seront utilisés. Ces deux domaines de recherche exploitent des techniques d'apprentissage (DL) pour modéliser la sémantique des textes et générer de nouvelles connaissances.

Plus précisément, nous avons montré dans un travail préalable [DGS+22] le potentiel des des approches " data-to-text " [PDL19a, RSSG20, PDL19b] pour la génération de réponses complexes.

Sous la co-tutelle de :

Notre objectif à long terme est de s'adapter au contexte de la recherche conversationnelle et de prendre en compte les interactions des utilisateurs et le contexte de la conversation [EPBG19, TY20], ainsi que d'inclure des caractéristiques orientées vers la tâche de recherche dans le processus de génération [FWZ+20, ZZW+20]. Deux lignes de recherche principales principales axes de recherche se distinguent :

- L'une est liée à la multiplicité des sources de données (texte, tableaux, figures, etc.) utilisées pour générer le texte et la structure de sortie.
- L'autre est plus liée à la satisfaction de l'utilisateur par rapport à la sortie en elle-même. Le document généré doit être à la fois complet, compréhensible et explicable.

L'application à des cas d'utilisation industrielle sera envisagée en collaboration avec l'équipe de développement d'Ecovadis.

Tous nos modèles seront évalués sur des benchmarks académiques, permettant une évaluation quantitative et la publication des résultats obtenus.

Profil recherché :

Master ou diplôme d'ingénieur en informatique ou en mathématiques appliquées en rapport avec l'apprentissage automatique, le traitement du langage naturel ou la recherche d'informations. Le candidat doit avoir un solide bagage scientifique, de bonnes compétences techniques en programmation, et doit pouvoir lire et écrire couramment l'anglais.

Début et durée (prévue) : Octobre/novembre 2022, 36 mois

Lien du sujet : <https://bit.ly/3KFtEWw>

Description du sujet (en anglais)

Context:

The perspective of new information retrieval (IR) systems (e.g., search-oriented conversational systems or systems supporting complex search tasks) has fostered research on theoretical IR models either leveraging or supporting users' interactions, for instance, through question clarification or interactive ranking models. However, very few works focus on the way of interacting with the user in natural language, which is critical for instance for conversational systems.

Project Description:

The main objective of the thesis is to design question-answering models aiming at solving multi-faceted information needs. More particularly, given a document collection, our goal is to generate structured and complete answers, covering all facets of a complex information need.

To do so, approaches and models from information retrieval (IR) and natural language processing (NLP) will be necessary. Both research fields exploit Deep Learning (DL) techniques to model semantics underlying texts and generate new knowledge. More precisely, we showed in a premise work [DGS+22] the potential of data-to-text approaches [PDL19a, RSSG20, PDL19b] for complex answer generation. Our long-term objective is to fit with the conversational search setting and to deal with users' interactions / conversational context [EPBG19, TY20] as well as include search task-oriented features in the generation process [FWZ+20, ZZW+20].

Sous la co-tutelle de :



INSTITUT DES SYSTEMES INTELLIGENTS ET DE ROBOTIQUE OFFRE DE THESE

Two main lines of research stand out:

- one is linked to the multiplicity of data sources (text, tables, figures, etc.) used to generate the output text and structure.
- another one is more linked to the user satisfaction regarding the output in itself. The generated document should both be complete, understandable and explainable.

Application to industrial use cases will be envisioned in collaboration with the development team at Ecovadis.

Required Profile:

Master or engineering degree in Computer Science or Applied Mathematics related to machine learning/natural language processing/information retrieval. The candidate should have a strong scientific background with good technical skills in programming, and be fluent in reading and writing English.

Starting and duration (expected): October/November 2022, 36 months

More info : <https://bit.ly/3KFtEWw>

Sous la co-tutelle de :

