Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

PhD thesis proposal

« Response generation models for solving multi-faceted information needs»

PhD proposal in Computer Science, main research fields : information science, information retrieval and access, natural language processing

Keywords: language generation, data-to-text, machine learning, deep learning

This position is proposed in collaboration between the ISIR laboratory (Sorbonne Université, Paris), the IRIT laboratory (Toulouse) and the Ecovadis company with a CIFRE contact.

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.

PHD objectives

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]. 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.

All our models will be evaluated on academic benchmarks, enabling quantitative evaluation and the publication of the obtained results.

Scientific supervisors:

The research work will be supervised by :

- Laure Soulier, Associate professor (ISIR laboratory, Sorbonne Université) - laure.soulier@isir.upmc.fr
- Karen Pinel-Sauvagnat, Associate professor (IRIT laboratory, Université Toulouse 3 - karen.sauvagnat@irit.fr
- Lynda Tamine, full professor (IRIT laboratory, Université Toulouse 3) - - lynda.tamine@irit.fr
- Sophia Katrenko, PhD, Ecovadis - skatrenko@ecovadis.com

Location :

The doctoral student may either be based in ISIR in Paris or in IRIT in Toulouse (preferably in Toulouse, but might be discussed with candidates). Regular exchanges will take place by video-conference as well as physical meetings.

ISIR laboratory :

The ISIR laboratory, located at the University Paris-Sorbonne, has slightly over 200 members. ISIR researchers work on the autonomy of machines and their ability to interact with human beings.

Research team = MLIA (Machine Learning and deep learning for Information Access) - https://www.isir.upmc.fr/equipes/mlia/presentation/

IRIT laboratory :

The IRIT laboratory is located in Toulouse and has over 600 members (it is one of the largest Joint Research Unit - UMR) at the national level). The laboratory has focused its research on five major scientific issues, one of them being the generation of intelligible information from raw data.

Research team: IRIS (Information Retrieval & Information Synthesis) - http://www.irit.fr/IRIS-site/

Ecovadis :
https://ecovadis.com
Ecovadis is the world's largest provider of business sustainability ratings. Since 2007, more than 90,000 compagnies have been rated. The Ecovadis team is composed of over 1000 professionals from 52 nationalies.

The Ecovadis methodology for sustainability rating is built on international sustainability standards, including the Global Reporting Initiative, the United Nations Global Compact, and the ISO 26000, covering 200+ spend categories and 160+ countries. Indicators cover four themes: (1) environment, (2) labor and human rights, (3) ethics and (4) Sustainable procurement.

Expected 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

How to apply ?

To apply, please email your application to supervisors. The application should consist of the following:

+ a curriculum vitae

+ transcript of marks according to M1-M2 profile or last 3 years of engineering school (with indication on the ranking if possible)

+ a covering letter

+ letter(s) of recommendation including at least one letter drawn up by a university referent

Interviews will be conducted as they arise and the position will be filled as soon as possible.

Bibliography

[DGS+22] Hanane Djeddal, Thomas Gerald, Laure Soulier, Karen Pinel-Sauvagnat, and Lynda Tamine. Does structure matter ? leveraging data-to-text generation for answering complex information needs. ECIR 2022 (short paper), to appear.

[DVRC17] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. Trec complex answer retrieval overview. TREC, 2017.

[EPBG19] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. Can you unpack that ? learning to rewrite questions-in-context. In Empirical Methods in Natural Language Processing, 2019.

[FWZ+20] Xiyan Fu, Jun Wang, Jinghan Zhang, Jinmao Wei, and Zhenglu Yang. Document summari- zation with vhtm : Variational hierarchical topic-aware mechanism. Proceedings of the AAAI Conference on Artificial Intelligence, 34 :7740–7747, Apr. 2020.

[PDL19a] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. pages 6908–6915. The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, 2019.

[PDL19b] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with entity modeling. pages 2023–2035. Proceedings of the 57th Conference of the Association for Compu- tational Linguistics, ACL 2019, 2019.

[RSSG20] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. A hierarchical model for data-to-text generation. 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I, volume 12035 of Lecture Notes in Computer Science, pages 65–80. Springer, 2020.

[TY20] Zhiwen Tang and Grace Hui Yang. Corpus-level end-to-end exploration for interactive sys- tems. Proceedings of the AAAI Conference on Artificial Intelligence, 34(03) :2527–2534, Apr. 2020.

[ZZW+20] Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, Ling Fan, and Zhe Wang. Topic-guided abstractive text summarization : a joint learning approach. 2020.