Leveraging data-to-text generation for answering complex information needs

Laboratory : IRIS team in the IRIT lab (Toulouse) or MLIA team at Sorbonne University (Paris)
Advisors :
— Karen Pinel-Sauvagnat (pinel@irit.fr)
— Laure Soulier (laure.soulier@lip6.fr)
— Lynda Tamine (tamine@irit.fr)

1 Contexte

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.

In this internship, we focus on the upstream part of the search process, once relevant documents have been identified in response to a complex information need, particularly characterized by multiple topical facets. Our motivation is to provide a complete and structured answer in natural language to the user on the top of the retrieval model. We, therefore, envision solving complex information needs with generative models (seq-to-seq models), particularly from the perspective of data-to-text generation [PDL19a, RSSG20, PDL19b]. This last category of models puts its attention on the notion of structure. One particular approach retained our attention [PDL19a] : a content selection and planning pipeline which aims at structuring the answer by generating intermediate plans.

This internship relies on a premise work [DGS+21] investigating the potential of data-to-text approaches for complex answer generation using the TREC CAR dataset [DVRC17]. With the long-term objective to fit with the conversational search setting, the objective of the proposed internship will be to extend this work in order to add :

— search task-oriented features [FWZ+20, ZZW+20]
— controllable features in the answer generation [SLS+18, PEM+21]
— users' interactions / conversational context [EPBG19, TY20]

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.

How to apply ? Send a CV, a motivation letter and Master records to advisors. Recommendation letters would be appreciated. Interviews will conducted as they arise and the position will be filled as soon as possible.

Références

[DGS+21] 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. *arXiv*, 2021.

[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 summarization 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 Computational Linguistics, ACL 2019, 2019.

[PEM+21] Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. A plug-and-play method for controlled text generation, 2021.

[RSSG20] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. A hierarchical model for data-to-text generation. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval - 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.

[SLS+18] Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text style transfer. *CoRR*, abs/1811.00552, 2018.

[TY20] Zhiwen Tang and Grace Hui Yang. Corpus-level end-to-end exploration for interactive systems. *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.