Identifying context for question-answering systems:
Use case: building data-to-text datasets for biological resources

Host team: MLIA team.
Encadrement: Laure Soulier (laure.soulier@lip6.fr) and Vincent Guigue (vincent.guigue@lip6.fr)
In collaboration with the "Muséum National d'Histoire Naturelle": Régine Vignes Lebbe and Thierry Bourgoin

Expected profile: Master or engineering degree in Computer Science or Applied Mathematics related to machine learning/natural language processing. 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 laure.soulier@lip6.fr and vincent.guigue@lip6.fr. Recommendation letters would be appreciated. Interviews will conducted as they arise and the position will be filled as soon as possible – the latest application date is set to 15th January.

Thesis perspective: The MLIA team will recruit at least one PhD. Student on the data-to-text generation topic for October 2022.

1 Contexte

Question-answering systems have attracted a lot of attention of the research community these last years and applications in industry are numerous, such as finding information in financial/medical documents or answering questions from customers. In the research community, the current task consists in answering a question on the basis of one or several documents (constituting the context of the question) [BCW14, RZLL16, YQZ+18, MZZH19]. Most approaches follow the "Retriever-reader" framework aiming at first identifying candidate documents/sentences willing to include the answer, and then finding the answer in this candidate. This last part can be carried out by either selecting a part of the sentence or generating a new sentence. This framework is generally performed sequentially but Lewis et al. [LPP+20, NO20, SWH+18] have also proposed an end-to-end trainable neural models in which both the document ranking and the answer selection functions are updated during the back-propagation.

In this internship, we propose to exploit question-answering systems for building new datasets for the data-to-text generation task. Data-to-text is a subfield of natural language generation and aims at transcribing structured data (tables, graphs, ...) into natural language descriptions [PDL19b, RSSG20, PDL19a, PDL19b, AKG+20]. However, this research field is limited by the small number of available datasets and their intrinsic peculiarities. Our objective here is to produce a new dataset based on biological domain gathering challenging properties. One peculiarity of biological datasets is that structured data and natural language summaries are not aligned, hindering the possibility of building supervision. We therefore propose to leverage question-answering systems to identify pairs of structured data-paragraphs. Structured data could be used for generating question-answer pairs (for instance with the Data-QuestEval framework [RSS+21]) and paragraphs could be seen as the context. The task would be thus reversed with an objective of finding the good paragraph given a question-answer pair.
Given the small size of available datasets, one challenge of such approach would be to learn the model with weak supervision or by leveraging domain adaptation techniques.

The intern objective would be thus to:

- Perform a literative review of question-answering models
- Proposing a model for context selection for academic datasets (as a proof of concept)
- Experimenting the approach on the biological resources
- Tackling the issue of the small amount of available data to train the model.

References


