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

[AKG+20] Oshin Agarwal, Mihir Kale, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Machine translation aided bilingual data-to-text generation and semantic parsing. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 125–130, Dublin, Ireland (Virtual), 12 2020. Association for Computational Linguistics.

[BCW14] Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 615–620. ACL, 2014.

[LPP+20] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[MZZH19] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. In *ACL*, 2019.

[NO20] Makoto Nakatsuji and Sohei Okui. Answer generation through unified memories over multiple passages. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3823–3829. ijcai.org, 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.

[RSS+21] Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. Data-questeval: A referenceless metric for data to text semantic evaluation. *CoRR*, abs/2104.07555, 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.

[RZLL16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

[SWH+18] Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. Leveraging context information for natural question generation. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 569–574. Association for Computational Linguistics, 2018.

[YQZ+18] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018.