Question Answering over Heterogeneous Graphs

Oana Balalau, Ioana Manolescu

September 2020

1 Context

In the SourcesSay project, we are interested in finding useful information in large datasets, to provide support for investigative journalism. Real-world events such as elections, public demonstrations, disclosures of illegal or surprising activities, etc. are mirrored in new data items being created and added to the global corpus of available information. Making sense of this wealth of data by providing a QA framework will facilitate the work of journalists. The team has an on-going collaboration with journalists from Le Monde.

The SourcesSay project builds upon the ConnectionLens framework that was developed in the team. The framework will ingest any type of dataset (text, CSV, JSON, XML, RDF, PDF, and relational database) and organized the information as a **heterogeneous graph**, where nodes represent important pieces of information (such as entities in text, nodes in RDF, non-null attributes in tuples in relational data and so on) and edges represent a structural or semantic connection between the nodes. Such connections are currently explored in our framework using keyword search. Specifically, given a set of $k >= 2$ search terms such as, e.g., "Assemblée Nationale" and "Russia", we find all the paths that lead from a node matching the former term to another that matches the second term. This enables finding that "Assemblée Nationale" and "Russia" are connected through the contract that the wife of a member of the Assemblée has with a Russian state-owned company. For a larger number of keywords, the query result will be a tree, where a keyword can be a node or an edge label in the tree.

While keyword search is a powerful tool, it requires that the user goes through all the answers (paths or trees) to find if the information of interest is there. For example, when using the two keywords "Assemblée Nationale" and "Russia", the user may in fact already know that his question really is "Which member of the Assemblée Nationale has financial interests in Russia?" Yet, the user may avoid adding keywords related to the financial domain, unsure which to chose.

The present project aims to devise a natural-language query answering module over ConnectionLens graphs. Such a module will be very valuable:

- To discover the content of a dataset, a user may ask aggregation questions, for example, how many entities exist in the dataset, how often is a certain entity mentioned, how many entities of a certain type exist (for example how many members of the Assemblée Nationale are present in our input dataset).
- To find interesting information, a user may ask precise questions, such as "Which member of the Assemblée Nationale has financial interests in Russia?" Note that while this question might be answered by a keyword search query (keywords "Assemblée Nationale" and "Russia"), however, looking for any kind of paths between two entities might find a lot of useless information, for example, the Assemblée Nationale is situated in Paris, which has the type capital like Moscow, which is situated in Russia.

2 Related Work

QA over a KB. There have been many approaches in this line of work, from crafting query templates that, once filled in, will be used to query the KB [UBL+12], to neural models, where the goal is to represent the question and the possible answers as latent vectors, where the correct answer should be

close in the embedding space to the question [BCW14]. Questions over KB are usually split in one-hop questions ("Who is the president of France?") and multi-hop questions ("To whom is the president of France married?"). As the name suggests, the difference is that the answer is situated either in the immediate neighborhood of the input entities (in a KB this may be encoded by the triple *Emmanuel Macron, presidentOf, France*) or at a few hops distance (we first find *Emmanuel Macron, presidentOf, France* and then *Brigitte Macron, spouseOf, Emmanuel Macron*). Multi-hop questions have increased the interest in KB completion, the task of adding missing triples to a KB [GML15].

QA over heterogeneous datasets. QA over several types of datasets has largely targeted QA over text and KBs. The challenge consists of enhancing the KBs with unstructured information so that more questions can be answered. Text has been used for adding to the KB triples obtained via OpenIE techniques [Mau16, FZE14], but also for ranking the possible answers of a question [SA16]. While research in QA over text and KBs has been driven by the huge increase in use of web search engines and personal assistants, QA over relational databases has been advertised as a tool of exploration of databases for non-technical users [VPS+19].

QA with complex questions. Complex questions are defined in literature as questions requiring set, logical, and arithmetic operations, for example "How many presidents did France have?" or "Who was the first president of France?" [Cha20]. Approaches for this task try to translate natural language in entities and constraints, for example by creating multi constraint query graphs [BDY+16], or translating a question into an executable program [ASK+19].

3 Thesis Objectives

There are many interesting directions to pursue in this thesis:

1. ConnectionLens graphs are not as structured as a curated, integrated knowledge base (such as DBPedia or Yago), but they are not completely unstructured (such as text). Our first goal is to take advantage of any existing structure to best answer questions over these datasets. This will entail trying different representations of the data, for example as proposed in [BMPS20], and/or trying to assign types to nodes, and normalize edge labels between nodes.

2. Our second goal is to be able to answer with high confidence a wide variety of types of questions, from simple one-hop questions to complex questions over ConnectionLens graphs. The requirement for high confidence comes from the importance of the application, as journalists will use our framework to improve or speed up their work. For this task, we will experiment with different neural models, in particular for representing our heterogeneous graphs, while keeping in mind recent criticism brought to these models [ASZ+20]. We will also tap into neural models proposed for answering complex questions [ASK+19] and the new direction of combining neural and symbolic computation [MBR+20].

3. Finally, our third goal is related to the need for transparency of sources under which journalists work. Any proposed solution for a QA system should be interpretable and open to scrutiny. Therefore, in addition to the answer to the question, the user will also have access to the explanation for the choice presented.

References

[ASK+19] Ghulam Ahmed Ansari, Amrita Saha, Vishwajeet Kumar, Mohan Bhambhani, Karthik Sankaranarayanan, and Soumen Chakrabarti. Neural program induction for kbqa without gold programs or query annotations. In *IJCAI*, pages 4890–4896, 2019.

[ASZ+20] Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. Realistic re-evaluation of knowledge graph completion methods: An experimental study.

In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1995–2010, 2020.

[BCW14] Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620, 2014.

[BDY+16] Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2503–2514, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[BMPS20] Irène Burger, Ioana Manolescu, Emmanuel Pietriga, and Fabian M. Suchanek. Toward visual interactive exploration of heterogeneous graphs. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020*, volume 2578 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[Cha20] Soumen Chakrabarti. Interpretable complex question answering. In *Proceedings of The Web Conference 2020*, pages 2455–2457, 2020.

[FZE14] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1156–1165, 2014.

[GML15] Kelvin Guu, John Miller, and Percy Liang. Traversing knowledge graphs in vector space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 318–327, 2015.

[Mau16] Mausam Mausam. Open information extraction systems and downstream applications. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, pages 4074–4077, 2016.

[MBR+20] Pasquale Minervini, Matko Bosnjak, Tim Rocktäschel, Sebastian Riedel, and Edward Grefenstette. Differentiable reasoning on large knowledge bases and natural language., 2020.

[SA16] Denis Savenkov and Eugene Agichtein. When a knowledge base is not enough: Question answering over knowledge bases with external text data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 235–244, 2016.

[UBL+12] Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. Template-based question answering over RDF data. In *Proceedings of the 21st international conference on World Wide Web*, pages 639–648, 2012.

[VPS+19] Ngoc Phuoc An Vo, Octavian Popescu, Vadim Sheinin, Elahe Khorasani, and Hangu Yeo. A natural language interface supporting complex logic questions for relational databases. In *International Conference on Applications of Natural Language to Information Systems*, pages 384–392. Springer, 2019.