Knowledge Graph Project Proposal

Alexander Prochnow

I. Significance

The scientific community benefits greatly from our modern connected world: Scientific papers are no longer distributed in physical copies, but can be accessed by anyone digitally the moment they are published. However, the information within these papers is mostly still isolated within text and figures, not taking advantage of our modern information technology.

An attempt towards more annotated, interconnected literature has been made by requiring authors to add hierarchical keywords to their papers, allowing for a simple, common categorization. These keywords, however, do not describe all aspects of a paper and are often too general for the niche subdomains that exist today. This makes literature research and getting a good grasp of a field harder, especially with the large volume of new literature published yearly.

To address these shortcomings, projects such as the Open Research Knowledge Graph (ORKG) offer a platform for authors to annotate important aspects of their papers by adding property-value pairs. However, these annotations must be made manually, and ORKG only offers a few simple suggestions based on a paper's abstract. Figure 1 shows an ORKG example that mostly includes these automated suggestions. This approach has two issues: The reliance on authors or contributors, as explained above, and the lack of a common vocabulary, which makes the construction of an interlinked Knowledge Graph (KG) difficult.

To solve these two issues, this project aims to assess the feasibility and quality of automatic knowledge graph creation from scientific papers, with the goal of linking papers using a shared open vocabulary such as Wikidata. Such a knowledge graph would allow researchers to get a more complete overview of their field, since it would not just include the papers annotated by authors on ORKG. Additionally, the KG would be much denser, since a common vocabulary would make the interconnectivity much greater.

II. Related work

Many papers in the field of automatic KG creation from text (Text-to-KG) focus on single or multiple modules within the Text-to-KG pipeline, e.g. Yan et al. [1] focus on entity linking, and Paolini et al. [2] focus on a joint entity and relation extraction. Some end-to-end approaches exist, e.g. the recent IBM Grapher [3]. However, this particular model does not link the extracted nodes and edges to concepts in a vocabulary.

Automatic Text-to-KG applied specifically to scientific papers can be seen by Sun et al. [4]. Here, only predefined entities relevant to social and behavioral sciences, such as sample size, were extracted. In contrast, I propose to use open vocabularies with many more entities and relations.

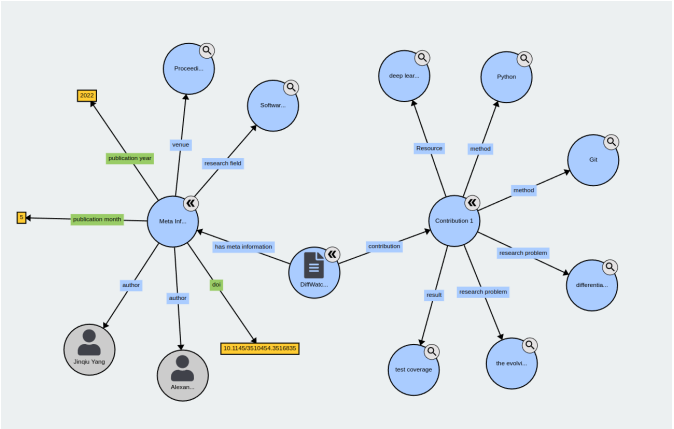

Fig. 1: Example of an ORKG paper knowledge graph (https://orkg.org/paper/R288464)

A project stemming from ORKG created a named entity extraction model specifically designed for computer science terminology (CS-NER[1]), making a step towards a common vocabulary in scientific paper knowledge graphs. However they limit themselves to seven extracted entity types, e.g. the research problem, solution or method. Additionally, CS-NER is only meant to be used on abstracts, while I aim to test a pipeline that uses the entire paper's text.

III. Goal and objectives

The overarching goal of this project is to create an easier and more complete way to do literature research, resulting in a better overview of a scientific field. For this, the state-of-the-art in automatic Text-to-KG will be investigated and a pipeline will be implemented to test the state-of-the-art models on scientific papers from within a specific scientific domain. The resulting KGs for each paper will first be evaluated individually, then as a single knowledge graph linking all individual paper KGs together.

IV. Methodology

Prior literature review has resulted in a first formulation of a model architecture for the Text-to-KG pipeline, containing three modules:

1) Preprocessing: Previous work [5] has shown that Coreference Resolution as a preprocessing step improves the Entity Extraction and Linking. For this I will use the spaCy extension NeuralCoref [6] from Hugging Face, which returns coreference scores between each pair of

[1]CS-NER: https://orkg-nlp-pypi.readthedocs.io/en/latest/services/services.html#cs-ner-computer-science-named-entity-recognition

words in a text. Using these scores, any references (e.g. pronominals) can be replaced by the entity they are refering to.
2) Entity Extraction and Linking: This technique can recognise entities in text and classify the entity type (person, location etc.) for disambiguation, then link to a corresponding public KG entry. One implementation of this is ReFinED [7] from Amazon Science, which links the entities it identifies to their Wikidata entries.
3) Relation Extraction and Linking: This module extracts the relations between entities and links them to properties found in public KGs. Here I will use REBEL from Babelscape [8], which links the relations it identifies to Wikidata properties.

A. Risks

In case one of the models within the implemented Text-to-KG pipeline performs significantly worse than the others, pulling the entire pipeline performance down, I will seek a replacement through further literature review. Another risk might be that scientific concepts might be underrepresented in Wikidata, since it is a general-purpose knowledge graph. In this case I will attempt to use a Text-to-KG method which can also identify entities and properties not found on Wikidata, such as the IBM Grapher [3], and compare the resulting KGs with the KGs created from the previous pipeline. In this way I may be able to identify the entities and relations missing on Wikidata and contribute. With this richer data, future work may then be able to fine-tune existing models specifically on a certain scientific domain.

V. Time plan and milestones

- Week of Feb. 27: Incorporate proposal feedback and implement first version of the proposed pipeline, testing on scientific papers during development. Milestone: Functioning Text-to-KG pipeline.
- Week of March 6: Evaluate first pipeline version on papers in multiple scientific fields. Make improvements to pipeline, tweak or exchange modules, then reevaluate. Milestones: KGs of scientific papers with evaluation and an improved pipeline.
- Week of March 13: Automatically link the created KGs, then evaluate interconnectivity and usefulness. Buffer for mentioned risks. Write report. Milestones: Interconnected KG of one or more scientific fields with evaluation and a first draft of a report.
- March 20-21: Finalize report, clean code repositories. Final deliverables: A repository with all code and KGs created as well as a project report.

VI. Anticipated results

I expect to produce a set of interconnected knowledge graphs, where each individual KG represents a scientific paper. These may contain sufficient information to be useful for certain scientific topics, but may be lacking quality for other subjects that are not as well recognized by the modules of the implemented Text-to-KG pipeline, or where the vocabulary is not represented well enough on public knowledge graphs. If the quality and usefulness is sufficient however, I may make the KGs available to others by hosting a SPARQL endpoint. Then I may continue this work after the project by devising a method for automatically extending and updating the KGs when new papers are published, as well as hosting a website with natural language search such that researchers can easily access the KGs.

References

[1] Yuchen Yan et al. “Dynamic Knowledge Graph Alignment”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.5 (2021), pp. 4564–4572. DOI: 10.1609/aaai.v35i5.16585. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16585.

[2] Giovanni Paolini et al. “Structured prediction as translation between augmented natural languages”. In: *ICLR 2021*. 2021. URL: https://www.amazon.science/publications/structured-prediction-as-translation-between-augmented-natural-languages.

[3] Payel Das Igor Melnyk Pierre Dognin. “Knowledge Graph Generation From Text”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*. 2022.

[4] Kexuan Sun et al. “Assessing Scientific Research Papers with Knowledge Graphs”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’22. Madrid, Spain: Association for Computing Machinery, 2022, pp. 2467–2472. ISBN: 9781450387323. DOI: 10.1145/3477495.3531879. URL: https://doi.org/10.1145/3477495.3531879.

[5] Tai Wang and Huan Li. “Coreference Resolution Improves Educational Knowledge Graph Construction”. In: *2020 IEEE International Conference on Knowledge Graph (ICKG)*. 2020, pp. 629–634. DOI: 10.1109/ICBK50248.2020.00094.

[6] Kenton Lee et al. “End-to-end Neural Coreference Resolution”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 188–197. DOI: 10.18653/v1/D17-1018. URL: https://aclanthology.org/D17-1018.

[7] Tom Ayoola et al. “ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking”. In: *NAACL 2022*. 2022. URL: https://www.amazon.science/publications/refined-an-efficient-zero-shot-capable-approach-to-end-to-end-entity-linking.

[8] Pere-Lluís Huguet Cabot and Roberto Navigli. “REBEL: Relation Extraction By End-to-end Language generation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204.