ASKG: An Approach to Enrich Scholarly Knowledge Graphs through Paper Decomposition with Deep Learning

Bowen Zhang, Sergio J. Rodríguez-Méndez and Pouya Ghiasnezhad Omran

Australian National University, Canberra ACT 2601, AU

Abstract

Knowledge Graphs (KGs) play a pivotal role in the field of artificial intelligence, yet the construction of such graphs often requires significant human resources. Automated KG construction technologies are key to achieving large-scale KGs construction. To address this, we have developed an automated Knowledge Graph Construction Pipeline (KGCP) and applied it to the enhancement of the Australian National University (ANU) Scholarly Knowledge Graph (ASKG), which comprehensively represents not only the metadata but also the scholarly knowledge encapsulated in the academic papers. This study introduces an innovative, automatic approach to KGs construction using an array of Natural Language Processing (NLP) techniques. Leveraging Named Entity Recognition (NER) models, key academic entities related to computer science are efficiently identified, such as Research Problems, Methods, Solution, Tool, Resource, Dataset, and Language. The ASKG is further enriched through Named Entity Linking (NEL) with Wikidata, keyword extraction, automatic summarisation, and the integration of entities from the Metadata Extractor & Loader and The NLP-NER Toolkit (MEL & TNNT).

Keywords

Knowledge Graph, Named Entity Recognition, Name Entity Linking, Deep Learning, Information Extraction, Knowledge Graph Construction

1. Introduction and Related Work

Academic KGs have been a focus in the field of cognitive intelligence. However, these KGs often concentrate on high-level metadata of papers, such as the author, date, venue, etc., while the in-depth exploration of paper content is often overlooked. This limitation hinders the full interpretation and utilisation of detailed knowledge within academic papers.

Addressing this issue is crucial as it can guide deeper analysis, identify emerging academic trends, reduce Large Language Models (LLMs) hallucination problem as well as enhance the training outcome of LLMs [1]. To tackle this, we implemented the PARSE (Papers And Relationships Semantic Extraction) component within our broader KGCP project, which decomposes

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece

Bowen.Zhang01@outlook.com (B. Zhang); Sergio.RodriguezMendez@anu.edu.au (S. J. Rodríguez-Méndez); P.G.Omran@anu.edu.au (P. G. Omran)

0000-0001-6045-8599 (B. Zhang); 0000-0001-7203-8399 (S. J. Rodríguez-Méndez); 0000-0002-4473-3877 (P. G. Omran)

academic papers and employs various NLP techniques and models for detailed knowledge extraction.

Numerous projects have been developed in the domain of academic KGs, such as AMiner [2], AceKG [2], and MAKG [3], which aggregate extensive information on researchers, publications, and citation relationships. However, their focus on fine-grained knowledge within papers is often insufficient.

ORKG [4] represents scholarly knowledge as structured data but it lacks detailed content analysis and fine-grained knowledge extraction. Other tools, like OpenAIRE [2] and ResearchRabbit [5], focus on promoting open academic exchange and offering functionalities like literature search and personalised summaries, visualisation, etc.

This paper presents an innovative approach to constructing KGs, emphasising the extraction of fine-grained knowledge from scholarly papers to enrich ASKG. Unlike the above-mentioned systems, our methodology entails section-wise parsing of academic papers adhering to the IMRaD (Introduction, Method, Results, and Discussion) structure. Many academic papers essentially adhere to the IMRaD structure. For those that do not follow the IMRaD format, we are in the process of implementing new tools and ontologies that can be customized according to the specific structure of each paper. We employ NLP techniques such as NER, NEL, automatic summarisation, and keyword extraction, individually applied to each IMRaD segment. This specific strategy distinctly positions ASKG from other KGs, offering a significant edge in gathering and processing academic data. Detailed comparisons with these platforms and tools can be found in our GitHub repository. Initial results suggest our method significantly enriches academic knowledge graphs, offering a more comprehensive and diverse data set, thus exemplifying the efficiency of our decomposition and refinement approach in knowledge graph construction.

2. KGCP Architecture: PARSE extension

Our ultimate goal is to expand the academic KGs by automatically extracting fine-grained knowledge through the structural decomposition of the documents (research papers). To achieve this, we are implementing and extending our KGCP[1] pipeline. As a key component of KGCP, the PARSE component is specifically focused on enriching ASKG by extracting meaningful knowledge from academic papers related to computer science.

As shown in Figure 1, firstly, we utilise web crawling to access ANU's target sources (academic web pages), MAKG, ScholarlyData, etc., automatically extracting information on researchers and their papers to build an academic paper dataset. We subsequently generate JSON files depicting paper metadata. PARSE operates in two primary phases. The first entails importing papers into the MEL & TNNT systems [6][7], extracting metadata, raw text, and general entities to enrich the existing ASKG.

In the second phase, PARSE extracts scientific field-specific knowledge from academic papers. Paper metadata described in JSON files is fed into a statistical analyser to obtain document set metadata and identify computer science papers. Using the targeted paper list, we send HTTP requests to the TNNT RESTful API, fetching the papers' original text. Then PARSE processes

[1]See: https://w3id.org/kgcp/, especially https://w3id.org/kgcp/PARSE

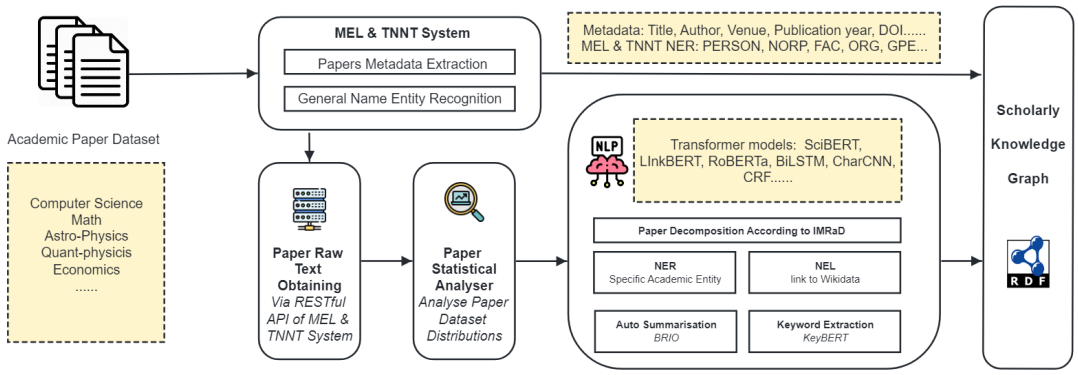

Figure 1: PARSE Structure

academic papers, segmenting them based on the IMRaD structure. We design and employ transformer-based NER models with RoBERTa, SciBERT, LinkBERT, etc. The text is sent to the NER module to identify computer science-related academic entities which are categorised as Research Problems, Methods, Solution, Tool, Resource, Dataset, and Language. Academic entities from the NER module are linked with Wikidata entities to enhance our knowledge graph. Meanwhile, we send different parts of the paper to the automatic summarisation model, BRIO, to generate summaries, and to the keyword model, KeyBERT, for keyword identification. All the outputs are processed to enrich the academic knowledge graph.

3. Evaluation, Discussion, and Current Work

Table 1
Comparison between the original and enriched ASKG

Metrics	Original ASKG	Enriched ASKG	Change (%)
Number of relation types	21	40	+90.48
Number of entity types	18	46	+155.56
Number of entities	235,314	1,215,106	+416.38
Number of triples	1,048,576	2,866,980	+173.42
Average degree	4.46	4.72	+5.83
Clustering coefficient	6.28e-05	1.86e-04	+196.18
Number of connected components	6	1	-83.33
Information density	4.46	2.36	-47.09

The comparison between the original and enriched ASKG, as shown in Table 1, reveals significant growth in aspects such as the number of relation types, entity types, entities, and triples, indicating enhanced structural diversity and information capacity. However, the information density has decreased, suggesting the enriched ASKG has become more sparse, posing a new research direction.

Listing 1 shows a portion of the output from PARSE. Unlike most traditional academic

knowledge graphs and previous ASKG, the enriched ASKG not only includes high-level abstract metadata such as authors and publication dates, but also contains more detailed academic information. This information includes, but is not limited to, keywords and a summary in each academic paper section, as well as more specific academic concepts like academic entities in each sentence and their locations in the academic paper.

```
@prefix askg-data: <https://www.anu.edu.au/data/scholarly/> .
@prefix askg-onto: <https://www.anu.edu.au/onto/scholarly#> .
@prefix domo: <https://www.anu.edu.au/onto/domo#> .
......

askg-data:Paper-5003681fa6a914 a askg-onto:Paper ;
    rdfs:label ''[SPICE: Semantic Propositional Image Caption Evaluation]-[Peter Anderson]-[2016]''@en ;
    askg-onto:hasSection askg-data:Abstract-1f35f04243f730 ,
        askg-data:Discussion-fc3bb8b300771b ,
        askg-data:Experiment-dc48c6d08186a7 ,
        ......
    askg-onto:paperLink ''http://arxiv.org/abs/1607.08822v1''^^xsd:string .

askg-data:Abstract-1f35f04243f730 a askg-onto:Abstract ;
    rdfs:label ''Paper-[SPICE: Semantic Propositional Image Caption Evaluation]-[Peter Anderson]-[2016] | Section-[
        Abstract]''@en ;
    domo:keyword askg-data:KeywordOfSection-0619f5fd0ab6a4 ,
    askg-onto:contains askg-data:Excerpt-ed1fc3dd5c08ab ,
    askg-onto:summary ''SPICE: Semantic Propositional Image Caption is a new automated caption evaluation metric
        ......''^^xsd:string .

askg-data:Excerpt-ed1fc3dd5c08ab rdfs:label ''Paper-['SPICE:␣Semantic␣Propositional␣Image␣Caption␣Evaluation'] |
    Section-['Abstract'] | Excerpt-[207]-[208]''@en ;
    askg-onto:inSentence ''there is considerable interest in the task of generating automatically image captions
        image captions [1,2]''^^xsd:string ;
    askg-onto:mentions askg-data:AcademicEntity-image_caption-Q39161486 ;
    askg-onto:wordIndexFrom ''207''^^xsd:int ;
    askg-onto:wordIndexTo ''208''^^xsd:int .

askg-data:AcademicEntity-image_caption-Q39161486 rdfs:label ''image caption''^^xsd:string ;
    owl:sameAs wd:Q39161486 ;
    skos:broader askg-onto:ResearchProblem .
......
```

Listing 1: Examples of PARSE output

With the enhanced ASKG, we propose a range of innovative use cases. One such use case is knowledge graph-based research trend analysis, illustrated in Table 2. While our study primarily focuses on capturing the dynamic evolution of academic research trends at the ANU, the methodology is designed to be adaptable and can be applied to other institutions as well. By executing SPARQL queries, we extract relevant data from the KGs and carry out a quantitative analysis, identifying the most mentioned academic entities and research problems, which can be interpreted as current research trends of the university's academic sources.

Rank	Research Problem	Frequency up to Jun. 2022	Frequency up to Dec. 2022	Rank Change
1	Optical Flow	230	260	+1 △
2	Modal Logic	231	258	-1 ▽
3	Image Captioning	144	180	+1 △
4	Blur Kernel	101	173	+1 △
5	Action Recognition	168	171	-2 ▽

Table 2
Example of Research Trend Analysis with ASKG

Research Trend Analysis can be used for academic performance management, resource

allocation, etc. It's worth noting that performing this level of refined analysis is challenging within traditional academic KGs that only include paper metadata. This is mainly because the metadata typically does not encompass in-depth descriptions of specific research problems or other academic knowledge, limiting our ability for a deep understanding of the dynamics within the research field. In contrast, our enriched ASKG can capture more information, thereby facilitating more detailed trend analysis.

Moreover, the enriched ASKG has a wider range of application scenarios, such as research relationship mining. By integrating diverse data including authors, research interests, academic entities, and summaries, it enables the discovery of overlooked patterns and potential cross-disciplinary collaborations between researchers through graph mining.

Currently, we continue applying the PARSE to other disciplines, such as astronomy and physics. Simultaneously, we are developing an innovative semantic query processing system (as an additional component of the KGCP) that combines LLMs with the enriched ASKG, aiming to improve the efficiency of academic information queries and the accuracy of context-based information retrieval from LLMs. In this system, user queries are translated into triple formats and then processed using SPARQL for graph matching, thereby supplying LLMs with more accurate and complete academic information.

We continue investigating and optimising the application of the LLMs and KGs in semantic searches and KG construction-related tasks, further advancing the fields of information retrieval and knowledge representation.

References

[1] F. Moiseev, Z. Dong, E. Alfonseca, M. Jaggi, SKILL: Structured Knowledge Infusion for Large Language Models, arXiv preprint arXiv:2205.08184 (2022).

[2] M. Nayyeri, G. M. Cil, S. Vahdati, F. Osborne, M. Rahman, S. Angioni, A. Salatino, D. R. Recupero, N. Vassilyeva, E. Motta, et al., Trans4E: Link prediction on scholarly knowledge graphs, Neurocomputing 461 (2021) 530–542.

[3] M. Färber, L. Ao, The Microsoft Academic Knowledge Graph enhanced: Author name disambiguation, publication classification, and embeddings, Quantitative Science Studies 3 (2022) 51–98.

[4] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246.

[5] R. Sharma, S. Gulati, A. Kaur, A. Sinhababu, R. Chakravarty, Research discovery and visualization using ResearchRabbit: A use case of AI in libraries, COLLNET Journal of Scientometrics and Information Management 16 (2022) 215–237.

[6] S. J. Rodríguez Méndez, P. G. Omran, A. Haller, K. Taylor, MEL: Metadata Extractor & Loader, in: ISWC (Posters/Demos/Industry), 2021.

[7] S. Seneviratne, S. J. Rodríguez Méndez, X. Zhang, P. G. Omran, K. Taylor, A. Haller, TNNT: The Named Entity Recognition Toolkit, in: Proceedings of the 11th on Knowledge Capture Conference, 2021, pp. 249–252.