Initial achievements in relation extraction from RNA-focused scientific papers

Emanuele Cavalleri[1], Mauricio Soto-Gomez[1], Ali Pashaeibarough[1], Dario Malchiodi[1], Harry Caufield[2], Justin Reese[2], Christopher J. Mungall[2], Peter N. Robinson[4], Elena Casiraghi[1,2,3], Giorgio Valentini[1,3] and Marco Mesiti[1,2,*]

[1]*Department of Computer Science, Università di Milano, Via Celoria 18, 20133 Milano*
[2]*Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[3]*ELLIS, European Laboratory for Learning and Intelligent Systems, Milan Unit, Italy*
[4]*Berlin Institute of Health - Charité, Universitätsmedizin, Berlin, 13353, Germany*

Abstract

Relation extraction from the scientific literature to comply with a domain ontology is a well-known problem in natural language processing and is particularly critical in precision medicine. The advent of large language models (LLMs) has paved the way for the development of new effective approaches to this problem, but the extracted relations can be affected by issues such as hallucination, which must be minimized. In this paper, we present the initial design and preliminary experimental validation of SPIREX, an extension of the SPIRES-based system for the extraction of RDF triples from scientific literature involving RNA molecules. Our system exploits schema constraints in the formulations of LLM prompts along with our RNA-based KG, RNA-KG, for evaluating the plausibility of the extracted triples. RNA-KG contains more than 9M edges representing different kinds of relationships in which RNA molecules can be involved. Initial experimental results on a controlled data set are quite encouraging.

Keywords

RNA-based Knowledge Graphs, relation discovery, LLM, Prompt Engineering, Link Prediction

1. Introduction

Ribonucleic acid (RNA) plays a critical role in the central dogma of molecular biology, serving as the intermediary between DNA and proteins, the building blocks of life. Beyond its traditional role in protein synthesis, RNA is involved in a variety of cellular processes, including gene regulation and catalysis, highlighting its importance in understanding the complexities of biological systems. RNA-KG [1] is the first ontology-based knowledge graph for representing

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] Corresponding author.

emanuele.cavalleri@unimi.it (E. Cavalleri); mauricio.soto@unimi.it (M. Soto-Gomez); ali.pashaeibarough@unimi.it (A. Pashaeibarough); dario.malchiodi@unimi.it (D. Malchiodi); jhc@lbl.gov (H. Caufield); justaddcoffee@gmail.com (J. Reese); CJMungall@lbl.gov (C. J. Mungall); peter.robinson@bih-charite.de (P. N. Robinson); elena.casiraghi@unimi.it (E. Casiraghi); giorgio.valentini@unimi.it (G. Valentini); marco.mesiti@unimi.it (M. Mesiti)

0000-0003-1973-5712 (E. Cavalleri); 0000-0001-5977-9467 (M. Soto-Gomez); 0000-0002-0559-1992 (A. Pashaeibarough); 0000-0002-7574-697X (D. Malchiodi); 0000-0001-5705-7831 (H. Caufield); 0000-0002-2170-2250 (J. Reese); 0000-0002-6601-2165 (C. J. Mungall); 0000-0002-0736-9199 (P. N. Robinson); 0000-0003-2024-7572 (E. Casiraghi); 0000-0002-5694-3919 (G. Valentini); 0000-0001-5701-0080 (M. Mesiti)

coding and non-coding RNA molecules and their interactions with other biomolecular data as well as with pathways, abnormal phenotypes and diseases to support the study and the discovery of the biological role of RNA. RNA-KG contains around 9M edges extracted from more than 50 public data sources and can be exploited to study RNA molecules and develop innovative graph algorithms to support knowledge discovery in data science.

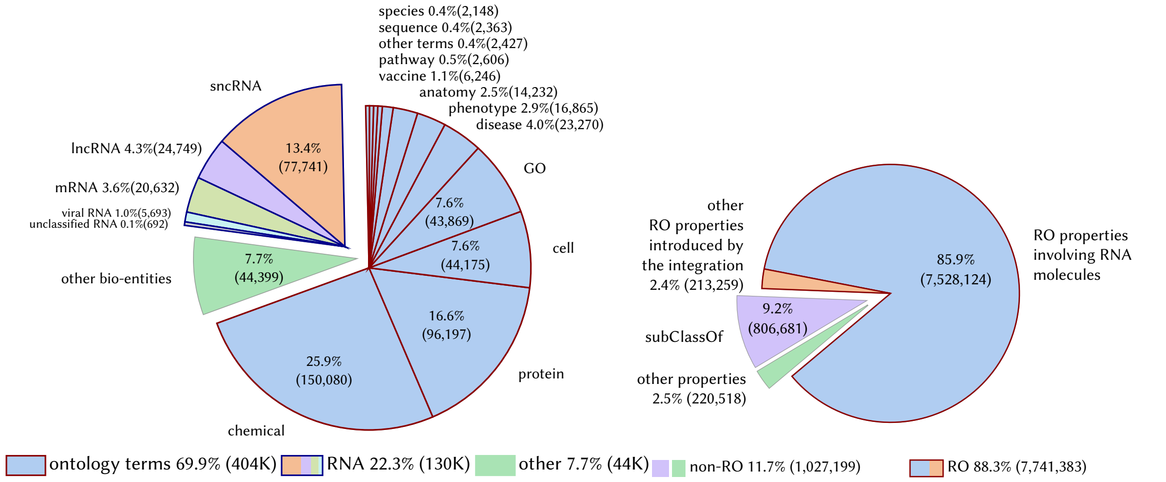
The manual ingestion of triples in a knowledge graph by expert curators is a time-consuming and costly operation and tools supporting them in the extraction of biological entities and their relationships from plain texts are highly demanding. The advent of LLMs [2] has paved the way for the development of new effective tools for this problem [3]. However, these techniques have shown different limitations, such as generating incorrect statements due to hallucinations (inaccurate, nonsensical, or irrelevant output in the given context) [4] and insensitivity to negations [5], that cannot be tolerated in sensitive domains like precision medicine. SPIRES (Structured Prompt Interrogation and Recursive Extraction of Semantics) [6] is a recently proposed knowledge extraction approach that exploits LLMs to identify instances of a knowledge schema expressed in terms of LinkML [7]. Since the schema contains a conceptualization of a given domain in terms of concepts, relationships, and properties we are interested in, it can be used for defining more effective LLM prompts. Additionally, SPIRES allows grounding of atomic textual elements as concepts taken from a variety of OBO Foundry ontologies [8].

Even if SPIRES has proven its efficiency in the extraction of triples from plain text according to bio-ontologies, there is the need to evaluate the reliability of the extracted triples both in terms of the generated identifiers (i.e. they correctly represent the identified entities) and the accuracy of their data source. In this paper, we address this problem by exploiting RNA-KG as a gold standard in the RNA world because it contains many interactions involving RNA molecules and can be used to evaluate the plausibility of the extracted triples.

To leverage SPIRES for its ability to extracting triples from texts and supporting experts in their validation, we present *SPIREX*, a system for the extraction of reliable triples from scientific papers. There are two main backbones of the system. On one side, SPIRES and the LinkML representation of the RNA-KG schema [9] allow the extraction of RDF triples compliant with the domain Ontology. On the other side, we use RNA-KG as a gold standard providing knowledge about interactions involving RNA molecules and use link prediction techniques to validate the 'plausibility' of the extracted triples; i.e., the likelihood of the triple to be part of RNA-KG. The initial experimental results on a manually curated testbed of 60 scientific texts are encouraging.

2. RNA-KG and SPIRES

RNA-KG [1] is the first knowledge graph encompassing biological knowledge about RNAs gathered from more than 50 public databases, integrating functional relationships with genes, proteins, chemicals, and ontologically grounded biomedical concepts. The current release of RNA-KG has a single component containing around 600K nodes and 9M edges and can be queried via SPARQL endpoint at https://RNA-KG.anacleto.di.unimi.it. Nodes are usually mapped to reference biomedical vocabularies and ontologies such as NCBI Gene Entrez identifiers for uniquely identifying genes and many kinds of non-coding RNAs (ncRNAs), Human Phenotype Ontology (HPO [10]) for phenotypes, Monarch merged disease ontology (Mondo [11]) for



(a) Distribution of nodes.

(b) Distribution of edges.

Figure 1: (a) node distribution according to node types; (b) edge distribution according to edge types.

diseases, and Gene Ontology (GO [12]) for annotating ncRNAs. Moreover, all the possible interactions are represented by means of the Relation Ontology (RO [13]). This ensures common semantics for the different relationships that can be extracted from the sources.

Figure 1a shows the distribution of nodes contained in RNA-KG (details in [1]). Nodes can be classified into nodes representing ontological terms and bio-entities lacking a direct mapping to ontological terms. Bio-entities have been further subdivided into RNA nodes, and non-RNA nodes (named `other bio-entities`) that contain, for instance, gene and nodes describing genomics features (e.g., nucleotide substitution). Figure 1b shows the distribution of edges in RNA-KG. Edges have been subdivided into three categories: i) edges representing RO properties that characterize interactions among RNA molecules in the considered sources; ii) other edges not belonging to RO properties; iii) edges representing the `subClassOf` relationships. The edges of the last two categories are introduced from the integration of bio-ontologies into RNA-KG and the lack of a dedicated ontology for RNA molecules. When RNA molecules cannot be precisely mapped to a reference ontology, they are included as `subClassOf` an appropriate class within Sequence Ontology (SO [14]).

SPIRES [6] is a recently proposed approach to information extraction that creates and refines prompts to maximize the effectiveness of LLMs by exploiting domain knowledge encapsulated through a schema expressed in LinkML [7]. By identifying and extracting relevant information from an input text, it adopts zero-shot or few-shot learning to identify and extract relevant entities and relationships among them, which are then normalized and grounded through ontologies and vocabularies. SPIRES is a general-purpose approach that can be used across a variety of domains and does not require specific training/tuning on the considered domain. SPIRES adopts an engineering approach for creating prompts for interacting with an LLM (like GPT3, GPT4) to improve the quality of the generated responses through the use of domain-specific schema. In this way, technical challenges for generative AI (e.g., constructing comprehensive real-world

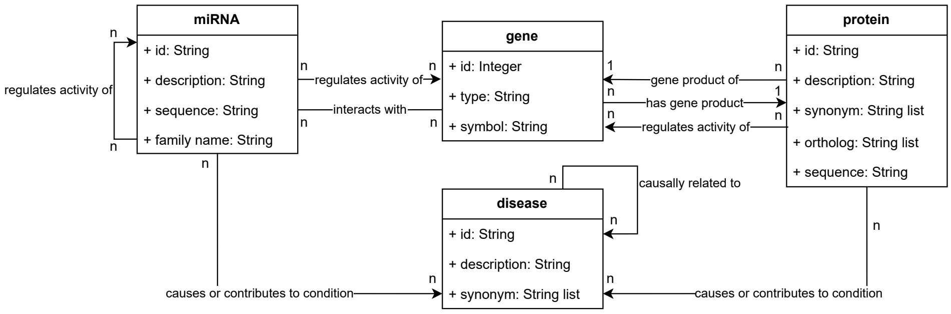


Figure 2: An excerpt of RNA-KG schema.

knowledge and improving the accuracy of automated responses) can be addressed.

The specification of this schema in LinkML contains the classes of entities and relationships among them within the specified domain. Classes can also include attributes (e.g., name, type, and list of synonyms) to enrich entity description. The LinkML schema is automatically processed to generate a list of prompts through which SPIRES interacts with a LLM. Each prompt of the list is submitted to the LLM for collecting information that is exploited for completing the following prompt by eventually considering the bio-ontologies (e.g., for changing a protein symbol with the corresponding identifier in an ontology). This recursive refinement process improves the quality of the information gathered through the LLM.

3. The SPIREX system

As shown in the architecture in Figure 3, SPIREX is composed of two modules: the SPIRES module is used for extracting the RDF triples from scientific abstracts. Then, an embedding of RNA-KG is used to validate the generated triples and score their level of plausibility.

SPIRES module for RNA-KG. Through the study of the scientific literature about RNA interactions, and the analysis of more than 50 data sources [1] all over the world, we have identified the kinds of relationships that can involve RNA molecules and reported them in a meta-graph [15]. Figure 2 shows an excerpt of the UML schema describing the entities that are connected to miRNA molecules through different kinds of relationships in the meta-graph.

Starting from its LinkML representation, a list of prompts specific for the RNA domain are generated according to which entities and the relationships contained in a text are extracted by considering the schema constraints. Moreover, SPIRES adopts bio-ontology of our domain (details in [9]) for producing source and target identifiers according to the RNA-KG identification scheme and RO predicates.

RNA-KG module for link prediction. The validation of new potential relations derived from the SPIRES module can be modeled as a link prediction task on RNA-KG, performed via either Graph Neural Networks (GNNs) or Random-Walk (RW) based methods for Graph Representation Learning. GNN approaches usually present scalability issues, while RW-based

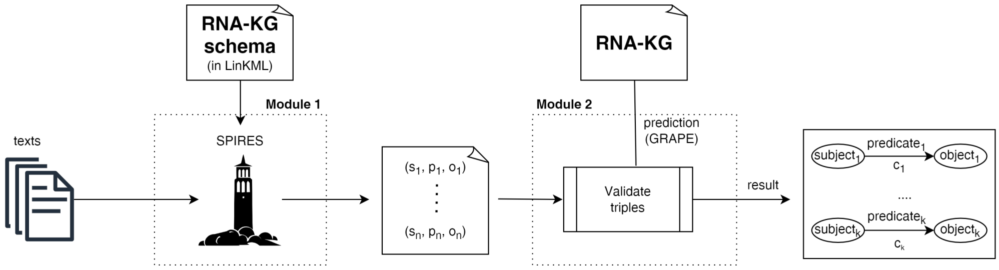


Figure 3: The SPIREX architecture.

graph embedding overcomes this problem by the use of random-walk approaches that sample the graph to construct a representation of the nodes (and edges) in a lower dimensional vector space that feeds traditional ML models.

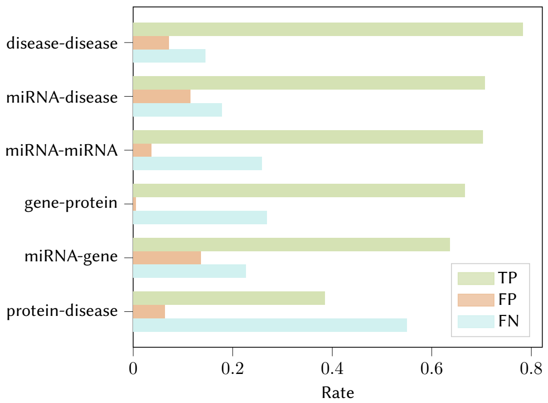
In SPIREX we have used Node2vec [16] for the embedding of RNA-KG. Node2vec is a well-known random walk-based approach that aims to capture the graph topology from the node neighborhoods. The model generates a set of second order random walks across the graph, that are used to train a shallow neural network to compute a vector representation of the graph components. One of the key features of Node2vec is the possibility to generate paths that focus either on the local or global structure of the graph, providing a great flexibility in the graph representation. Our system uses the implementation available in GRAPE [17], a software resource specifically designed for the manipulation and embedding of large graphs.

4. Preliminary experimental results

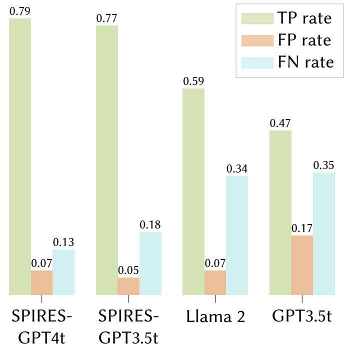
Experiments have been realized for both modules of SPIREX. For the first module, we evaluated the prediction accuracy of SPIRES in extracting triples in a set of manually annotated documents. We also compared SPIRES with base LLMs to verify the advantage of using LinkML in the specification of the domain schema. For the second module, we checked if the simple predictive model can generate reasonable scores on RNA-KG. Finally, we assessed the ability of the predictor to evaluate the plausibility of triples extracted through SPIRES according to RNA-KG.

SPIRES prediction accuracy and comparison with base LLMs. As described in [9], a corpus of 60 scientific articles related to RNA molecules and their interactions has been gathered from PubMed, ResearchGate, and Google Scholar. Starting from them, we have identified abstracts, discussions, or specific subsections within the domain of interest. They have been manually annotated with the entities and the six kinds of interactions that can be extracted from them (reported in the y-axis of the diagram in Figure 4a).

For evaluating the predictions, we have used standard metrics (precision, recall, and F-score) by considering the True Positive (TP), False Positive (FP), and False Negative (FN) according to the manually tagged paragraphs. As shown in Figure 4a, the obtained results, using GPT3.5-turbo in SPIRES for each category of interaction, indicate a consistent trend where TP rate tends to be higher with respect to both FP and FN rates. The only exception is for protein-



	TP	FP	FN	F-score	Precision	Recall
Total	304	41	134	0.76	0.88	0.69



	F-score	Precision	Recall
SPIRES-GPT4t	0.88	0.91	0.86
SPIRES-GPT3.5t	0.86	0.94	0.81
Llama 2	0.74	0.89	0.64
GPT3.5t	0.64	0.73	0.57

(a) TP, FP, and FN results on 60 texts.

(b) Comparing SPIRES, Llama, GPT on 20 texts.

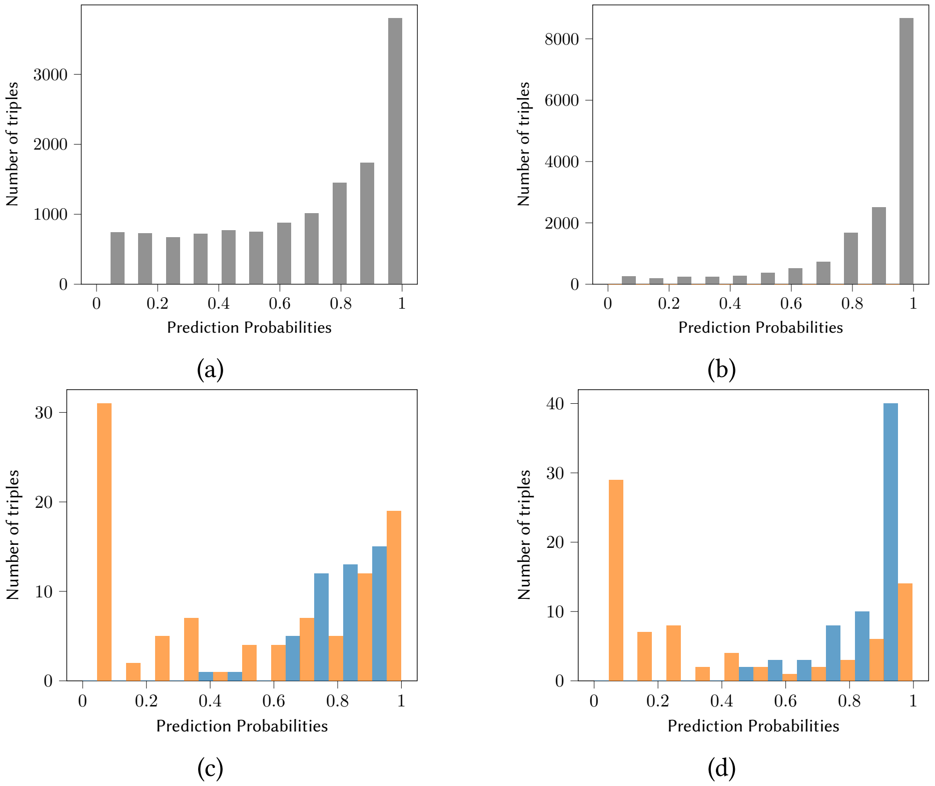
Figure 4: Evaluation of SPIRES on relation extraction involving protein, miRNA, disease, and gene entities and comparison against different LLMs.

disease relations, where FN rate is higher than TP rate. We noticed that many protein-disease relations are undetected, often because they are expressed in complex ways and this can lead to inaccurate entity recognition. Despite this, the overall precision remains remarkably high and, in biomedicine, this is preferable because it prioritizes certainty over ambiguity.

We have assessed the performance of SPIRES by considering as baseline approaches OpenAI GPT (ver. GPT3.5-turbo) and Llama 2 [18] (ver. llama-2-70b-chat). As back-end LLM of SPIRES, we have considered both GPT3.5-turbo and GPT4-turbo. We have manually grounded instances and relationships that can be extracted from 20 documents among those considered in the previous experiment. Regarding the prompt to be used with the base LLM system, we have considered a simple one requesting to extract triples from the considered text with an explicit request for mapping the extracted concepts to appropriate terminologies. Given that both OpenAI GPT and Llama 2 caution that the ontology identifiers provided are hypothetical and might not align with actual identifiers in the ontologies, and considering the general community advice against relying on IDs from an LLM [19], we decided to substitute the grounding process with our manually curated look-up tables [1].

As shown in Figure 4b, SPIRES outperforms baseline LLMs used alone both in terms of precision and recall. The histogram points out a high increment in TP rate and a decrease in FP and FN rates when adopting SPIRES for extracting relations that adhere to a specified schema within texts. Furthermore, when adopting GPT4-turbo in SPIRES the recall metric improves due to the lower FN rate with a positive effect on the F-score.

Evaluation of the plausibility of SPIREX predictions. For evaluating the plausibility, a restricted RNA-KG view has been considered that roughly corresponds to the schema in Figure 2 focusing on the predictions of miRNA-disease relationships. More precisely, we have considered two different settings of the hold-out procedure to evaluate prediction performance.



(a)

(b)

(c)

(d)

Figure 5: Distribution of the miRNA-disease edge predictions of the RNA-KG link prediction module. Node2vec predictions: (a) on the test set of RNA-KG$_{\Delta_{disease}}$; (b) on the test set of RNA-KG$_{\Delta_{10\%}}$; (c) of the edges predicted by SPIRES with RNA-KG$_{\Delta_{disease}}$; (d) of the edges predicted by SPIRES with RNA-KG$_{\Delta_{10\%}}$. In c)-d), 'blue'/'orange' bars represent triples 'included'/'not included' in the view.

In the first one, named RNA-KG$_{\Delta_{disease}}$, the test set corresponds to triples involving miRNAs and diseases from the source RNAdisease [20], while training set corresponds to the remaining miRNA-disease triples of RNA-KG; in the second one, named RNA-KG$_{\Delta_{10\%}}$, we randomly included in the test set 10% of miRNA-disease triples, and in the training set the remaining 90%, independent of their original source, guaranteeing to maintain the graph connectivity, according to a connected Monte-Carlo hold-out strategy [17].

We directly applied node2vec to the prediction of miRNA-disease edges according to the RNA-KG$_{\Delta_{disease}}$ and RNA-KG$_{\Delta_{10\%}}$ experimental settings, using a Multi-Layer-Perceptron trained on the node2vec edge embeddings. The default parameters adopted in GRAPE have been chosen. Figure 5a and 5b show that the triples in the test set exhibit high probabilities, in both experimental settings. As reported in Figure 5a, with RNA-KG$_{\Delta_{disease}}$, $\sim$63% of these triples are associated with a score higher than 0.6. In the case of RNA-KG$_{\Delta_{10\%}}$, we notice that $\sim$88% of the test set was correctly classified with a probability higher than 0.6 and $\sim$74% with a probability higher than 0.8 (see Figure 5b). Node2vec is thus a reasonable predictor of miRNA-disease edges to be included in the RNA-KG and can be used to assess the plausibility of SPIRES predictions.

To assess the ability of node2vec in evaluating the plausibility of the triples extracted by

SPIRES, we have considered true positive triples extracted from our manually curated dataset involving miRNAs and diseases. Figures 5c and 5d show the distribution of the probabilities predicted by node2vec on the miRNA-disease edges extracted by SPIRES. Specifically, blue columns represent the number of miRNA-disease triples that are already included in RNA-KG, whereas orange columns represent the number of triples that are missing in RNA-KG. In both cases, node2vec is able to correctly classify almost all the tuples already present in the partial KG but can also discriminate between plausible and implausible new triples, offering a potential validation tool. Indeed in both experimental settings, we can identify a set of edges included in RNA-KG that are predicted with a high probability by both SPIRES and node2vec (blue bars), but also a set of edges extracted by SPIRES and predicted with a high probability by node2vec, even if these edges are not present in RNA-KG (orange bars). These last edges can be considered as possible new candidates for miRNA-disease relationships. In Figures 5c and 5d, the orange bars denote relationships identified by SPIRES, yet assigned a low probability by node2vec. These edges can be considered “uncertain” in the sense that they are not confirmed by an independent edge prediction method that exploits the topological characteristics of the RNA-KG. We believe these results can be improved by considering expanded views of RNA-KG and more complex ML methods capable of accommodating its inherent heterogeneity.

5. Concluding remarks

In this paper we have described the initial steps in the design and development of the SPIREX system for the extraction of meaningful triples from scientific papers that exploit RNA-KG as a gold standard for checking the plausibility of the extracted triples. The initial experimental results are encouraging of the effectiveness of the proposed tool. At the current stage, we have used a basic link prediction measure for assessing the relationship’s plausibility according to the knowledge graph’s current state. However, a much more accurate measure should be developed that takes into account other factors (like the number of times the relationship has been identified in different sources, the presence of the relationship in other sources of information, or the coherence of the relationship with respect to the other triples extracted from the same scientific paper). We are also considering the adoption of other link prediction methodologies, especially those for heterogeneous graphs that can easily scale with big KGs. Finally, even if the approach has been tested in the context of RNA-KG, we would like to generalize it to other application domains that exploit biomedical KGs (e.g. [21]) for extracting new facts from texts.

Datasets. Experiments have been realized by using the following datasets: (schema and docs) https://doi.org/10.5281/zenodo.10671796; (RNA-KG): https://doi.org/10.5281/zenodo.10078876.

Acknowledgements. This research was in part supported by the “National Center for Gene Therapy and Drugs based on RNA Technology”, PNRR-NextGeneration EU program [G43C22001320007] and in part by the MUSA - Multilayered Urban Sustainability Action - Project, funded by the PNRR-Next-Generation EU program ([G43C22001370007], Code ECS00000037).

References

[1] E. Cavalleri, et al., RNA-KG: An ontology-based knowledge graph for representing interactions involving RNA molecules, 2023. arXiv:2312.00183.

[2] R. Bommasani, et al., On the opportunities and risks of foundation models, 2021. arXiv:2108.07258.

[3] A. J. Thirunavukarasu, et al., Large language models in medicine, Nature Medicine 29 (2023) 1930–1940. doi:10.1038/s41591-023-02448-8.

[4] Z. Ji, et al., Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. doi:10.1145/3571730.

[5] A. Ettinger, What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models, Transactions of the Association for Computational Linguistics 8 (2020) 34–48. doi:10.1162/tacl_a_00298.

[6] J. H. Caufield, et al., Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning, Bioinformatics 40 (2024) btae104. doi:10.1093/bioinformatics/btae104.

[7] S. Moxon, et al., The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics, in: International Conference on Biomedical Ontologies, 2021, pp. 148–151. URL: https://ceur-ws.org/Vol-3073/paper24.pdf.

[8] R. Jackson, et al., OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies, Database 2021 (2021). doi:10.1093/database/baab069.

[9] E. Cavalleri, M. Mesiti, On the extraction of meaningful RNA interactions from scientific publications through LLMs and SPIRES., In: 8th Int'l workshop on Data Analytics solutions for Real-LIfe APplications., 2024.

[10] P. N. Robinson, et al., The human phenotype ontology: A tool for annotating and analyzing human hereditary disease, The American Journal of Human Genetics 83 (2008) 610–615. doi:10.1016/j.ajhg.2008.09.017.

[11] N. A. Vasilevsky, et al., Mondo: Unifying diseases for the world, by the world, 2022. doi:10.1101/2022.04.13.22273750.

[12] M. Ashburner, et al., Gene ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25–29. doi:10.1038/75556.

[13] C. Mungall, et al., oborel/obo-relations: 2023-08-18 release, 2023. doi:10.5281/zenodo.8263469.

[14] K. Eilbeck, et al., The sequence ontology: a tool for the unification of genome annotations, Genome Biology 6 (2005). doi:10.1186/gb-2005-6-5-r44.

[15] E. Cavalleri, et al., A meta-graph for the construction of an rna-centered knowledge graph, in: Bioinformatics and Biomedical Engineering, Springer Nature Switzerland, Cham, 2023, pp. 165–180. doi:10.1007/978-3-031-34953-9_13.

[16] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, 2016. arXiv:1607.00653.

[17] L. Cappelletti, et al., GRAPE for fast and scalable graph processing and random-walk-based embedding, Nature Computational Science 3 (2023) 552–568. doi:10.1038/s43588-023-00465-8.

[18] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023.

[19] T. Groza, H. Caufield, D. Gration, G. Baynam, M. A. Haendel, P. N. Robinson, C. J. Mungall, J. T. Reese, An evaluation of GPT models for phenotype concept recognition, 2023. arXiv:2309.17169.

[20] J. Chen, et al., RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction, Nucleic Acids Research 51 (2022) D1397–D1404. doi:10.1093/nar/gkac814.

[21] T.J. Callahan, et al., An Open-Source Knowledge Graph Ecosystem for the Life Sciences, Scientific Data 11(1) (2024) 363. doi:s41597-024-03171-w.