Literature-based Hypothesis Generation: Predicting the evolution of scientific literature to support scientists

Tarek R. Besold[a], **Uchenna Akujuobi**[a], **Samy Badreddine**[a], **Jihun Choi**[a], **Hatem Elshazly**[a], **Frederick Gifford**[a], **Chrysa Iliopoulou**[a], **Kana Maruyama**[a], **Kae Nagano**[a], **Pablo Sanchez**[a], **Thiviyan Thanapalasingam**[a], **Alessandra Toniato**[a], **Christoph Wehner**[a]

[a] *Sony AI,* tarek.besold@sony.com, uchenna.akujuobi@sony.com, samy.badreddine@sony.com, jihun.a.choi@sony.com, hatem.elshazly@sony.com, frederick.gifford@sony.com, chrysa.iliopoulou@sony.com, kana.maruyama@sony.com, kae.nagano@sony.com, pablo.sanchez2@sony.com, alessandra.toniato@sony.com

1. Introduction

Science is advancing at an increasingly quick pace, as evidenced, for instance, by the exponential growth in the number of published research articles per year [1]. On the one hand, this poses an increasingly pressing challenge: Effectively navigating this ever-growing body of knowledge is tedious and time-consuming in the best of cases, and more often than not becomes infeasible for individual scientists. On the other hand, from an AI point of view, scientific literature offers a great opportunity: The body of published research works offers a vast collection of highest-quality—literally expert-reviewed—data about the relationships of concepts and the governing laws of our physical world.

Making use of the opportunity in order to mitigate the challenge, computational systems have been introduced which aim to support human researchers in the initial phase of the scientific process by automatically extracting hypotheses from the knowledge contained in published resources, i.e., by performing automated hypothesis-generation (HG). Famously, [2] systematically used a scientific literature database to find potential connections between previously disjoint bodies of research, as a result hypothesizing a (later confirmed) curative relationship between dietary *fish oils* and *Raynaud's syndrome*. Swanson and Smalheiser then automatized the search and linking process in the ARROWSMITH system [3]. Their work and other more recent examples such as [4, 5, 6] demonstrate the usefulness of computational methods in extracting latent information from the vast body of scientific publications.

In the following, we summarize the current state of our efforts to contribute to the development of a fit-for-use HG system. Our report includes recent developments regarding approaches to literature-based HG and work aiming to make the resulting type of HG system fit-for-use by scientists through the provision of flanking explanatory information.

2. Hypothesis generation via temporally-aware link prediction

In the context of HG, where the goal is to predict novel relationships between entities extracted from scientific publications, comprehending prior relationships is of paramount importance. Similar to [7, 8, 9], we believe that modeling the temporal evolution of these relationships thus plays a critical role in constructing an effective and resilient HG model.

To harness the temporal dynamics, we frame the HG task as a temporal graph problem. Given a sequence of timestamped graphs $G = \{G_0, G_1, ..., G_T\}$, each graph at time t is represented as $G_t = (V_t, E_t, F_t)$, where V_t is the set of entities (nodes, domain concepts), E_t is the set of relations (edge types), and $F_t \subseteq V_t \times E_t \times V_t$ are factual connections. These sets change over time, reflecting the graph's dynamic nature. The objective is to predict which previously unconnected nodes in G_T should be linked.

In [10], we introduced *THiGER*, a batch contrastive temporal node-pair embedding method for link prediction. THiGER uses a hierarchical transformer framework to effectively capture and learn from temporal information inherent in the input knowledge graphs, enabling efficient parallel temporal information aggregation. We also introduced *THiGER-A*, an incremental training approach incorporating an active curriculum learning strategy that trains the model on high-utility samples whilst also mitigating label bias arising from unobserved connections.

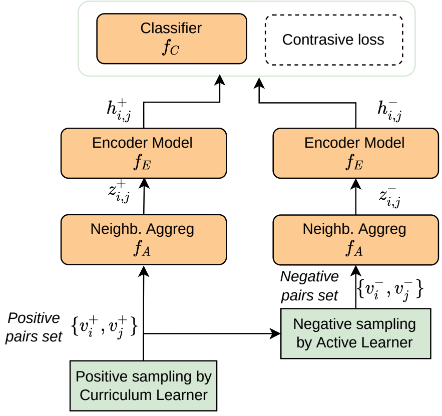

Fig. 1: General overview of *THiGER-A*, highlighting the Active Curriculum Learning strategy. [10]

Figure 1 shows the whole THiGER(-A) model. Let $v_i, v_j \in V_T$ be nodes denoting two concepts. The pair is assigned a positive label $y_{i,j} = 1$ if a corresponding edge (i.e., a link) is observed in G_T. That is, $y_{i,j} = 1$ iff $e(v_i, v_j) \in E^T$, otherwise 0. The model predicts a score $p_{i,j}$ that reflects $y_{i,j}$.

This link prediction score is given by a neural classifier $p_{i,j} = f_C(h_{i,j}^{[0,T]}; \theta_C)$, where $h_{i,j}^{[0,T]}$ is an embedding vector for the node pair, calculated using a hierarchical transformer encoder (see Figure 2). The input to the hierarchical encoder layer is the independent local node pair embedding aggregation at each time step. Subsequently, the local node pair embeddings aggregation is processed by the aggregation layer. At each hierarchical layer, temporal node pair embeddings are calculated for a sub-window of fixed size. The entire encoder architecture is denoted as $f_E = f_E^l : l = 1 \ldots L$.

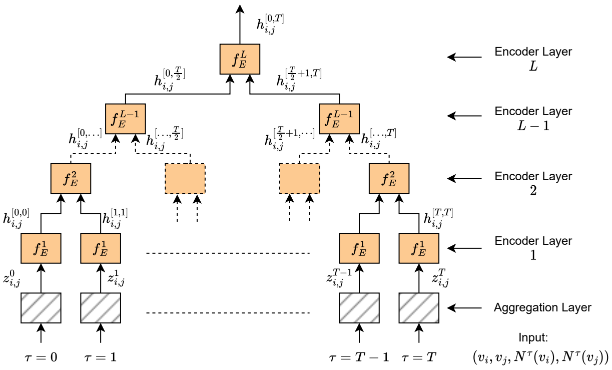

Fig. 2: Temporal Hierarchical Multilayer Encoder taking as input the aggregated node pair embeddings at each time step, outputting the generalized node pair embedding across all time steps. [10]

In [10] we modeled the classifier $f_C(.; \theta_C)$ using a multilayer perceptron network (MLP), the neighborhood aggregation $f_A(.; \theta_A)$ is based on GraphSAGE [11], and $f_E(.; \theta_E)$ uses the described multilayer transformer encoder network. Empirical validation of our approach showed it to be effective in established temporal-graph benchmarks and on real-world datasets from the food-health domain.

3. More than just hypotheses

If the aim is to make HG systems truly useable and useful for researchers in a scientific domain, it is not sufficient to output only an isolated novel hypothesis. Instead, additional information helping the domain expert to explain and contextualize the predicted hypotheses is strictly necessary. We, thus, are working to complement our HG system with additional modules performing this explanation task.

3.1 KGExplainer

Working towards providing the information scientists require to make sense of the generated hypotheses, in [12] we introduced *KGExplainer*, a post-hoc and local explainable AI method tailored for Knowledge Graph Embedding (KGE) models. It directly decodes the latent representations created by KGE models, encoding the graph's topology and the interactions between its entities. KGExplainer leverages the principle that similar embeddings reflect similar behaviors within the symbolic structure of the Knowledge Graph. By analyzing the subgraph neighborhoods of these entities, symbolic regularities in the form of conjunctive clauses are discovered. KGExplainer translates these regularities into symbolic rules, or triples comprehensible to humans, thereby uncovering the rationale behind the models' predictions in local subgraph contexts.

KGExplainer uses five steps to compile the explanations: After getting the k-nearest neighbors in the latent space of a predicted triple (Step 1), it creates positive and negative entity-pairs from the nearest neighbors (Step 2), and mines all possible clauses and their frequency within the subgraph neighborhood of the pairs (Step 3). It then identifies the most descriptive clauses for positive entity-pairs with the help of a surrogate model (Step 4), and finally grounds the most descriptive clauses to create an explanation (Step 5). In doing so, KGExplainer provides immediate, faithful explanations without retraining, in doing so facilitating real-time application on large-scale knowledge graphs.

3.2 GenACOX

True to the observation that different stakeholders and end-users have varying preferences for how explanations are presented [13, 14, 15], in addition to *KGExplainer* we most recently also explored generative, instance-based post-hoc methods that identify the minimal and most relevant set of facts that maximally influence a KGE prediction. We refer to this class of models as the Generative Ant Colony Optimization Explainer (*GenACOX*), which has several desired properties. First, the generative nature of *GenACOX* enables to find explanations that may include previously unknown facts. Second, since the space of possible facts is huge—of size $|V|^2 \times |E|$—*GenACOX* leverages ant colony optimization (ACO) [16] to efficiently explore this vast search space. Third, *GenACOX* uses a fast-inference surrogate model that approximates the KGE scoring function. This surrogate allows for i) quick evaluation of candidate fact sets as potential explanations; ii) reduces time and hardware requirements; and iii) enable applicability to black-box KGE models. That is, to generate explanations even in scenarios where the KGE model is proprietary so we can query it for predictions but lack access to its internal mechanisms or training pipeline.

4. First steps on a long way

What we presented are only pieces of a larger puzzle. Making HG systems maximally useful for scientists will require further exploration of the role these systems can play in actual day-to-day research work—a process that crucially needs to involve scientists throughout for input and feedback.

Acknowledgments

We want to thank Jesus de la Fuente, Pedro Giesteira Cotovio, Angel Hsing-Chi Hwang, Priyadarshini Kumari, Alessandro Lonardi, Susana Nunes, and Cliona O'Doherty for their input and support at different stages of the described project.

References

[1] Karen White. Publications Output: U.S. Trends and International Comparisons | NSF - National Science Foundation — ncses.nsf.gov. https://ncses.nsf.gov/pubs/nsb20214, 2021. [Accessed 14-March-2025].

[2] Don R Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986.

[3] Don R Swanson and Neil R Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, 91(2):183–203, 1997.

[4] Jung-wei Fan and Yves A Lussier. Word-of-mouth innovation: hypothesis generation for supplement repurposing based on consumer reviews. In *AMIA Annual Symposium Proceedings*, volume 2017, page 689. American Medical Informatics Association, 2017.

[5] Aaron Trautman. *Nutritive Knowledge Based Discovery: Enhancing Precision Nutrition Hypothesis Generation*. PhD thesis, The University of North Carolina at Charlotte, 2022.

[6] Agustín Borrego, Danilo Dessì, Daniel Ayala, Inma Hernández, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, David Ruiz, and Enrico Motta. Research hypothesis generation over scientific knowledge graphs. *Knowledge-Based Systems*, 315:113280, 2025.

[7] Uchenna Akujuobi, Michael Spranger, Sucheendra K Palaniappan, and Xiangliang Zhang. T-pair: Temporal node-pair embedding for automatic biomedical hypothesis generation. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2988–3001, 2020.

[8] Uchenna Akujuobi, Jun Chen, Mohamed Elhoseiny, Michael Spranger, and Xiangliang Zhang. Temporal positive-unlabeled learning for biomedical hypothesis generation via risk estimation. *Advances in Neural Information Processing Systems*, 33:4597–4609, 2020.

[9] Huiwei Zhou, Haibin Jiang, Weihong Yao, and Xun Du. Learning temporal difference embeddings for biomedical hypothesis generation. *Bioinformatics*, 38(23):5253–5261, 2022.

[10] Uchenna Akujuobi, Priyadarshini Kumari, Jihun Choi, Samy Badreddine, Kana Maruyama, Sucheendra K Palaniappan, and Tarek R Besold. Link prediction for hypothesis generation: an active curriculum learning infused temporal graph-based approach. *Artificial Intelligence Review*, 57(9):244, 2024.

[11] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[12] Christoph Wehner, Chrysa Iliopoulou, and Tarek R. Besold. From latent to lucid: Transforming knowledge graph embeddings into interpretable structures, 2024.

[13] Anahid N. Jalali, Bernhard Haslhofer, Simone Kriglstein, and Andreas Rauber. Predictability and comprehensibility in post-hoc xai methods: A user-centered analysis. *ArXiv*, abs/2309.11987, 2023.

[14] Andrea Rossi, Donatella Firmani, Paolo Merialdo, and Tommaso Teofili. Explaining link prediction systems based on knowledge graph embeddings. *Proceedings of the 2022 International Conference on Management of Data*, 2022.

[15] Ngoc Luyen Lê, Marie-Hélène Abel, and Philippe Gouspillou. Combining embedding-based and semantic-based models for post-hoc explanations in recommender systems. *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4619–4624, 2023.

[16] M. Dorigo. Optimization, learning and natural algorithms. *Ph.D. Thesis, Politecnico di Milano*, 1992.