Classifying Scientific Topic Relationships with SciBERT

Alessia Pisu[1,*], Livio Pompianu[1], Angelo Salatino[2], Francesco Osborne[2,3], Daniele Riboni[1], Enrico Motta[2] and Diego Reforgiato Recupero[1]

[1]*Department of Mathematics and Computer Science, University of Cagliari, IT*

[2]*Knowledge Media Institute, The Open University, UK*

[3]*Department of Business and Law, University of Milano Bicocca, IT*

Abstract

Current AI systems, including smart search engines and recommendation systems tools for streamlining literature reviews, and interactive question-answering platforms, are becoming indispensable for researchers to navigate and understand the vast landscape of scientific knowledge. Taxonomies and ontologies of research topics are key to this process, but manually creating them is costly and often leads to outdated results. This poster paper shows the use of SciBERT model to automatically generate research topic ontologies. Our model excels at identifying semantic relationships between research topics, outperforming traditional methods. This approach promises to streamline the creation of accurate and up-to-date ontologies, enhancing the effectiveness of AI tools for researchers.

Keywords

Research Topics, Ontology Generation, Language Models, Knowledge Graph Generation, SciBERT

1. Introduction

The current generation of AI technologies, such as smart search engines, recommendation systems, and question-answering applications, significantly aids researchers in exploring and interpreting scientific literature [1]. Despite this, the rapid growth of scientific publications, increasing by about 2.5 million papers annually [2], poses a substantial challenge. Although large language models have revolutionised natural language processing (NLP) [3], they still encounter limitations to process extensive text volumes and understand the broader context of a research area.

To address this, scientific knowledge graphs (SKGs) [4], such as SemOpenAlex[1], AIDA-KG[2], ORKG[3], CS-KG[4], became increasingly popular, providing structured and formal representations of research publications.

Research topics are essential for describing research concepts within SKGs, making ontologies of research topics (e.g., MeSH, UMLS, CSO, NLM) crucial for organising and querying academic information [5]. Altogether, they empower intelligent systems to efficiently navigate and

Joint Proc. of Posters, Demos, Workshops, and Tutorials of the 20th Int.l Conf. on Semantic Systems (SEMANTiCS 2024)

[*] Corresponding author.

[1] SemOpenAlex - https://semopenalex.org

[2] AIDA-KG - https://w3id.org/aida

[3] ORKG - https://orkg.org/

[4] CS-KG - https://w3id.org/cskg

understand academic literature, including advanced search engines, interactive conversational agents, analytics dashboards, and academic recommender systems.

However, manually creating ontologies of research topics is costly and time-consuming, often resulting in outdated representations. To address this challenge, several approaches have been proposed, including the integration of ontology learning with crowdsourcing methods, combining statistical analysis with user feedback [6], or utilising citation-based clustering of research papers to infer research topics from the titles and abstracts of documents within clusters [7]. Another approach is Klink-2 [8], which produced the Computer Science Ontology (CSO) [9], a widely adopted resource with about 14K topics and 159K semantic relationships.

In the same direction, this poster paper explores the use of SciBERT for generating research topic ontologies. Our goal is to develop a method that incorporates language model technology to update CSO and construct large-scale ontologies across scientific disciplines. We developed a model to identify four semantic relationships (*supertopic*, *subtopic*, *same-as*, and *other*) between research topics and compared its performance to traditional feature-based solutions. Preliminary results show that the transformer-based model significantly outperforms traditional models. The gold standard and code are available on a GitHub repository[5].

2. Materials and Methods

In this section, at first we describe the addressed task and the used datasets. Then, we illustrate a traditional feature-based approach, and our transformer-based technique.

2.1. Task Definition and Datasets

In this work, we address a single-label multi-class classification problem. The task is to classify the relationship between a pair of research topics (t_A, t_B) according to four categories which are essential for ontology generation:

- *supertopic*: t_A is a parent topic of t_B. E.g., *ontological languages* is a broader area than *owl*
- *subtopic*: t_A is a child topic of t_B. E.g., *nosql* is a specific area within *databases*
- *same-as*: t_A and t_B are different labels for the same concept. E.g., *haptic interface* and *haptic device*
- *other*: t_A and t_B do not relate according to the above categories. E.g., *blockchain* and *user interfaces*

In this context, *other* can refer to either negative samples or alternative semantic relationships not currently considered by our method, such as *partOf*, or *contributesTo*.

For our gold standard, we selected portions of the Computer Science Ontology [9] that have been manually checked and improved. CSO is a large ontology covering 14K research topics, providing an extensive and fine-grained representation of Computer Science. It was automatically generated using the Klink-2 algorithm [8] on 16 million scientific articles.

CSO comprises four primary semantic relationships. Among them, *superTopicOf* and *relatedEquivalent* essentially correspond to our *superTopic* and *same-as* relationships, respectively. To

[5] Gold standard and code - https://github.com/aleessiap/LeveragingLMforGeneratingOntologies.git

construct the gold standard, we selected 4,713 *superTopicOf* triples from the CSO and designated them as *superTopic* instances. Additionally, we chose 3,034 *relatedEquivalent* triples to represent equivalence using the *same-as* relation. We also derived 4,713 *subTopic* relationships by reversing the *superTopic* relationships. Lastly, we randomly paired topics to create 5,151 *other* relationships, ensuring that none of these pairs shared any of the previously identified relationships within the CSO. The resulting gold standard dataset consists of 17,611 triples, divided into 15,154 triples (86%) for the training set, 2,166 triples (12.3%) for the validation set, and 291 triples (1.7%) for the test set. To prevent bias, we ensured that topic pairs in one set do not appear in another. Moreover, each test set triple includes at least one topic not present in the training set. These measures make the test set more challenging compared to those used for Klink-2 [8]. In order to compute features involving the linkage of topics to relevant papers used in our feature-based method, we queried AIDA-KG [10], a KG considering 25 million publications linked to research topics in CSO.

2.2. Feature-based Method

Our classification task is commonly approached exploiting numerical features, usually measuring the frequency and common usage of the two topics [8]. Extracted feature vectors are then classified through mathematical functions or machine learning algorithms [8]. We devised a feature-based classification method using the following features for each pair of topics (t_A, t_B):

- *occA*: the frequency of t_A appearing in paper abstracts
- *occB*: the frequency of t_B appearing in paper abstracts
- *cooccurrenceAB*: the frequency of both t_A and t_B appearing together in abstracts
- *subsumption*: the degree of overlap between the co-occurring topics, computed as $subsumption = \frac{cooccurrenceAB}{occA} - \frac{cooccurrenceAB}{occB}$

The first two features indicate the popularity of a topic. The third feature quantifies the relatedness of two topics. The fourth feature assesses the hierarchical relationship between the topics. After normalising the features, we trained two ensemble machine learning models: Gradient Boosting (GB) and Random Forest (RF); varying the number of estimators from 10 to 3000 to determine the optimal configuration.

2.3. Language Model-based Method

Our method leveraging language models relies on SciBERT [11], an extension of BERT [12], which is a highly regarded model for its ability to effectively understand and process human language. SciBERT, trained on scientific literature from Semantic Scholar, enhances BERT's capabilities by focusing on the scientific domain.

To address our classification task we fine-tuned SciBERT using the training set described in Section 2.1. Specifically, we used the *scibert-scivocab-uncased* model from *Huggingface*. As optimiser, we selected *AdamW* [13] to prevent overfitting in large models. For the fine-tuning process, we provided the model with the surface forms of the two topics, separated by a semicolon. For each couple of topics, we also provided the correct relationship class from the training set. We experimented with varying the number of epochs from 1 to 10, maintaining 50 warm-up steps. Our best-performing model was achieved when training for five epochs.

Table 1
Experimental results. GB = Gradient Boosting, RF = Random Forest.

Classifier		Feature-based GB	Feature-based RF	Lang. Model-based
Accuracy		0.5842	0.6426	**0.9141**
Precision	*supertopic*	0.5424	0.5634	**0.9143**
	subtopic	0.4815	0.6200	**0.9452**
	same-as	0.5167	0.5804	**0.9615**
	other	0.8621	**0.8793**	0.8286
	average	0.6007	0.6608	**0.9124**
Recall	*supertopic*	0.4211	0.5263	**0.8421**
	subtopic	0.3421	0.4079	**0.9079**
	same-as	0.7750	0.8125	**0.9375**
	other	0.8475	0.8644	**0.9831**
	average	0.5964	0.6528	**0.9177**
F-score	*supertopic*	0.4740	0.5442	**0.8767**
	subtopic	0.4000	0.4921	**0.9262**
	same-as	0.6200	0.6771	**0.9494**
	other	0.8547	0.8718	**0.8992**
	average	0.5872	0.6463	**0.9129**

3. Evaluation

Using the test set described in Section 2.1, we evaluated the three methods outlined in the previous section: *Gradient Boosting* and *Random Forest* (both feature-based), and *SciBERT* (language model-based). We compared their performance using accuracy, precision, recall, and F-score, which are standard metrics for text classification.

Table 1 reports the experimental results. The language model-based method was far superior to the feature-based methods in all areas, achieving an impressive F1 score of 0.9129. This was over 27% higher than the other methods. Among the feature-based approaches, Random Forest performed better. The language model-based method was particularly effective in recognising *superTopic* and *subTopic* relations, where feature-based methods struggled, likely due to the presence of unfamiliar topics in the test set.

The language model-based method generally priorities precision over recall, particularly for the relations *superTopic*, *subTopic*, and *same-as*. However, for the *other* relation, it tends to miss some semantic connections, resulting in lower precision compared to recall. This suggests the model may incorrectly classify some related topics as *other*, an issue we intend to explore further in future research.

4. Conclusions

In this poster paper, we introduced a new method based on SciBERT to identify the relationship between research topics and conducted a comparative analysis against feature-based solutions. We fine-tuned a SciBERT model using a gold standard of triples derived from CSO. The model achieved an F1 score of 0.9129, a 27% improvement over methods using numerical features. These findings are significant given the growing demand for detailed ontologies to enhance content characterization in scientific KGs

In our future work, we aim to develop an innovative method for creating taxonomies of research topics to improve CSO and create large-scale ontologies across different scientific fields. We plan to combine language models and numerical features using knowledge injection techniques and experiment with recent large language models. We also intend to explore potential challenges when applying these techniques to other research domains and assess the impact of cross-disciplinary applications.

References

[1] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial intelligence for literature reviews: Opportunities and challenges, arXiv preprint arXiv:2402.08565 (2024).

[2] L. Bornmann, R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, Journal of the Association for Information Science and Technology 66 (2015) 2215–2222. doi:10.1002/asi.23329.

[3] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, PLoS digital health 2 (2023) e0000198.

[4] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artificial Intelligence Review (2023) 1–32.

[5] A. Salatino, T. Aggarwal, A. Mannocci, F. Osborne, E. Motta, A survey on knowledge organization systems of research fields: Resources and challenges, 2024. URL: https://arxiv.org/abs/2409.04432. arXiv:2409.04432.

[6] G. Wohlgenannt, A. Weichselbraun, A. Scharl, M. Sabou, Dynamic integration of multiple evidence sources for ontology learning, Journal of Information and Data Management 3 (2012) 243–254.

[7] OpenAlex, Openalex: End-to-end process for topic classification, 2024. URL: https://docs.google.com/document/d/1bDopkhuGieQ4F8gGNj7sEc8WSE8mvLZS/edit.

[8] F. Osborne, E. Motta, Klink-2: Integrating multiple web sources to generate semantic topic networks, in: The Semantic Web - ISWC 2015, Springer International Publishing, Cham, 2015, pp. 408–424.

[9] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, E. Motta, The computer science ontology: a large-scale taxonomy of research areas, in: The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17, Springer, 2018, pp. 187–205.

[10] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, Aida: A knowledge graph about research dynamics in academia and industry, Quantitative Science Studies 2 (2021) 1356–1398.

[11] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, 2019. arXiv:1903.10676.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[13] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.