Building Knowledge Graphs in Heliophysics and Astrophysics

Fech Scen Khoo[1(✉)], Megan Mark[2], Roelien C. Timmer[3],
Marcella Scoczynski Ribeiro Martins[4], Emily Foshee[5], Kaylin Bugbee[6],
Gregory Renard[7], and Anamaria Berea[8]

[1] University of Oldenburg, 26111 Oldenburg, Germany
fech.scen.khoo@uni-oldenburg.de
[2] Florida Institute of Technology, Melbourne, FL 32901, USA
[3] University of New South Wales, Sydney, NSW 2052, Australia
[4] Federal University of Technology, Ponta Grossa, PR 84017-220, Brazil
marcella@utfpr.edu.br
[5] University of Alabama, Huntsville, AL 35805, USA
elf0005@uah.edu
[6] NASA Marshall Space Flight Center, Huntsville, AL 35808, USA
kaylin.m.bugbee@nasa.gov
[7] The Applied AI Company (AAICO), Redwood City, CA 94062, USA
[8] George Mason University, Fairfax, VA 22030, USA
aberea@gmu.edu

Abstract. We propose a method to build a fully connected knowledge graph in the scientific domains of heliophysics and astrophysics, using word embeddings from BERT which are adaptively fine-tuned to these domains. We extract the scientific concepts automatically by a keyword extractor. The graph nodes representing these concepts are connected and weighed based on the cosine similarities computed from their fine-tuned embeddings. Our method is able to capture various meaningful scientific connections, and it incorporates the possibility to enable knowledge discovery.

Keywords: Knowledge graph · Knowledge discovery · Language models

1 Introduction

As we carry out our daily scientific work within our own expertise, the ever expanding knowledge web inevitably becomes unstructured in a way that it can be disconnected among concepts, let alone if there is a meaningful relation hidden between different domains. Within and across domains, the same scientific term can mean differently to the respective communities, for instance the term *radiation*: it could be related to X-ray observation for astronomers, or UV radiation

E. Métais et al. (Eds.): NLDB 2023, LNCS 13913, pp. 1–14, 2023.
https://doi.org/10.1007/978-3-031-35320-8_15

effect to the skin for biologists. On the other hand, cross-disciplinary researches have become increasingly important, as scientific discoveries in modern days usually benefit from huge collaborative efforts.

Therefore, structuring the knowledge base can encourage more dialogues and make possible new discoveries across the domains. Most of all, it will save the researchers' time in sorting out the ingredients they need for their research, and efforts can be well invested into the thought processes, experiments and etc., shedding lights on numerous questions or mitigations from how does the Sun affect lives on earth, to what is the origin of the Universe, and so on.

In science, to accept or rule out a concept or theory it requires sufficient experiments, observations and etc., which take time. With this in mind, how could we build a knowledge graph (KG) that incorporates probable or even accelerates new discoveries as well? In this regard, we turn into the investigation of how strongly or distantly connected are the given concepts.

On the other hand, from the perspective of data, the challenge adds up as we do not have a labeled dataset to train on for a Named Entity Recognition (NER) task, nor a ground truth to validate against. Apart from that, we also do not have an ontology which is typically used as a foundation to build a knowledge graph. Therefore, to extract the relevant entities, we will be using an automatic keyword extractor, and we rely on human experts in our team for validation. A naive strategy for us would be to start from constructing a smaller-sized knowledge graph that can be well verified in the process.

In our approach, we mine the texts using a controlled set of terms. We will specify these controlled terms in our experiments. The extracted keywords from the texts are taken to be related, and are examined in more detail where they are found to carry higher cosine similarities of at least 0.5. As a result, we will present specifically several pairs of the scientific terms or concepts that we obtained to illuminate what our embedding-based method using the fine-tuned language models can accomplish. In particular, we work on the domains of heliophysics and astrophysics, as we have the related knowledge expertise, and these 2 domains are known to be closely connected.

Our contributions from the present work are:

- We created 4 fine-tuned language models in heliophysics and in astrophysics: helioBERT, hierarchical helio-astroBERT, large helioBERT, and large hierarchical helio-astroBERT.
- We propose a novel method to build a knowledge graph based on cosine similarities for heliophysics and astrophysics domains.

2 Related Work

Some recent techniques such as logical reasoning and post-processing operations have been applied as refinement methods for knowledge graphs (e.g. automatic completion and error detection) [1]. Reasoning can deal with automatically deriving proofs for theorems, and for uncovering contradictions in a set of axioms. It is widely adopted in the Semantic Web community, leading to the development of a larger number of ontology reasoners. For example, if a person is defined to be the capital of a state, this is a contradiction, since cities and persons are disjoint, i.e., no entity can be a city and a person at the same time. Some approaches for knowledge graph building have implemented reasoning when new axioms are to be added (NELL dataset, PROSPERA).

To validate a KG, a naive but popular approach is to randomly sample triples from the KG to annotate manually. A triple is considered correct if the corresponding relationship is consistent with the domain expertise [2], hence the KG accuracy can be defined as the percentage of triples in the KG being correct – a sampling approach. While in terms of the quality of the extracted entities themselves, the most common approach is again human evaluation [3]. In general, the target entities can be extracted by a (fine-tuned) NER model through e.g. flairNLP[1]. Also, the flairNLP text embedding library comes with its word embeddings Flair, and options such as GloVe [4] and BERT [5], or a combination of different embeddings can be chosen. On the other hand, one can consider some structure graph metrics such as the Structure Hamming Distance metric [6], which can be applied to compare the built KG with a known ontology as a reference graph.

There has been an increased interest in generating knowledge graphs for heliophysics and astrophysics domains, such as the NASA Heliophysics KNOWledge Network project [7] and other initiatives [8,9]. We draw our inspiration from a recent approach that proposes a language model for astronomy and astrophysics known as astroBERT [10]. The language model astroBERT is found to outperform BERT on NER task on the data of astronomical content. Similarly, another BERT variant, SciBERT [11] which was proposed earlier, has been fine-tuned on scientific texts with 18% from computer science and 82% from biomedical domains.

3 Methodology

Our methodology consists of two main components. It begins with data collection, followed by a knowledge graph construction which requires a fine-tuning procedure of the language model BERT. We propose the following machine learning pipeline (Fig. 1):

First, to enrich our primary dataset from NASA's Science Mission Directorate (SMD), we collected abstracts from arXiv from the year 2021. These abstracts are more descriptive than SMD. They widen the knowledge spectrum in our SMD dataset, as they contain research findings from a larger scientific community. Thus this data addition potentially provides multiple (new) connections between knowledge entities. From this pool of abstracts from various research fields that cover for instance *Astrophysics*, and *Condensed Matter*, to further

[1] https://github.com/flairNLP/flair.

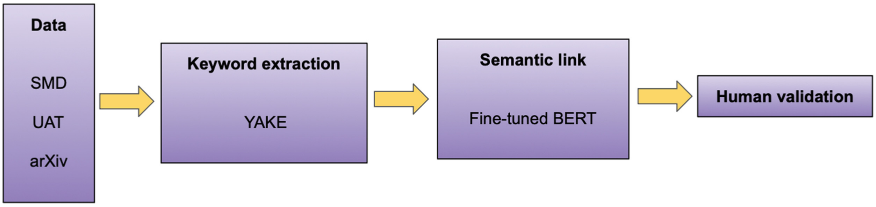

Fig. 1. Pipeline for building a knowledge graph.

align with the scientific context in our primary SMD dataset, we will narrow down our experiments to some specific research areas. For this, we define a number of different sets of scientific terms using two principal sources: SMD, and additionally Unified Astronomy Thesaurus (UAT), to extract only the relevant abstracts, by simply requiring that the set of terms (or any of them) form part of the abstract. With this, we have indirectly established to a certain degree a document similarity among the selected texts before we proceed to building the respective knowledge graph.

Next, we choose to extract only tri-grams out of our text data using YAKE [12]. Tri-gram turns out to be an optimal choice as it is sufficient to account for scientific terms such as *James Webb Space* (*James Webb Space Telescope* in full), and *Schwarzschild black holes* (a bi-gram *black holes* in general). YAKE which stands for Yet Another Keyword Extractor is an automatic keyword extractor, which includes readily the text pre-processing procedure that involves tokenization and stopword removal. The algorithm is succeeded by a statistical feature extraction then evaluated for a term score, followed by an n-gram (tri-gram in our case) keyword generation where its score is built out of the term score. The final step of the YAKE algorithm consists of data deduplication which further improves the ranking of the relevant keywords. For our purpose, we have chosen to extract a total of 20 keywords per text, ranked by the distance similarity metric, that is the Sequence Matcher (SEQM), implemented in YAKE. The lower the SEQM is, the more relevant or important the associated keyword is. These extracted keywords will serve as the nodes in our knowledge graph. These nodes are linked directly, that is, considered connected since these keywords come from a same pool of texts extracted using a particular set of terms (which represent a certain research topic) as explained before.

Now the question remains on how related the nodes are. We evaluate the strength of the connection or the semantic link between the nodes by computing cosine similarities based on the fine-tuned word embeddings from BERT. The threshold for cosine similarity value varies in each of our experiment, ranging from a minimum of 0.5 to 0.8 (1 being the highest), where below the set minimum value, we regard the pairs of entities as not strongly connected, and hence discard them for further analysis. We use the pre-trained transformer-based language model, BERT (BERTbase), and adaptively fine-tune the model on our datasets in order to shift BERT into our knowledge domains. Using the fine-tuned embeddings, we obtain a representation of our knowledge graph. For visualization, we use a package, igraph to create the graphs.

To effectively utilize the word embeddings, we propose the following 4 variations of BERT shown in Fig. 2, fine-tuned on a Masked Language Modeling (MLM) task using our datasets.[2] In an MLM task, a portion of the words in a sequence is masked and BERT learns to predict the masked tokens based on the context.

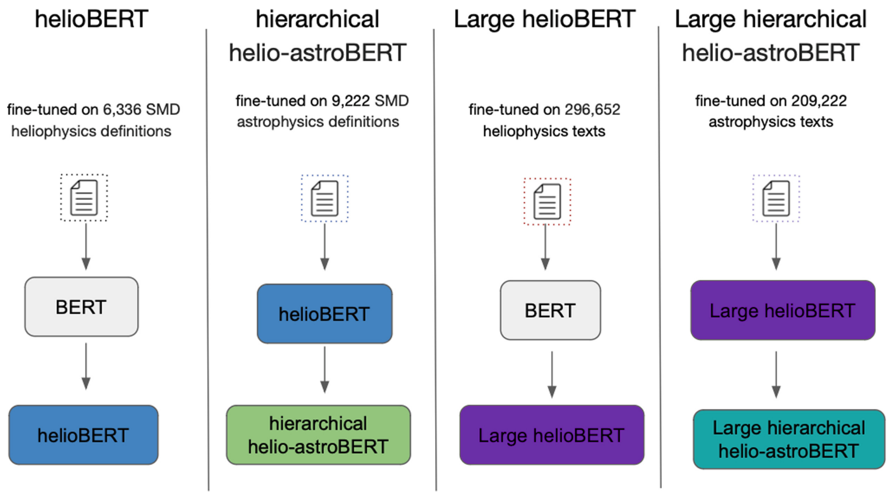
helioBERT
fine-tuned on 6,336 SMD heliophysics definitions
BERT
helioBERT
hierarchical helio-astroBERT
fine-tuned on 9,222 SMD astrophysics definitions
helioBERT
hierarchical helio-astroBERT
Large helioBERT
fine-tuned on 296,652 heliophysics texts
BERT
Large helioBERT
Large hierarchical helio-astroBERT
fine-tuned on 209,222 astrophysics texts
Large helioBERT
Large hierarchical helio-astroBERT

Fig. 2. BERT variations, differently fine-tuned on heliophysics and astrophysics texts.

Model I. helioBERT: BERT was trained on heliophysics texts.
Model II. Hierarchical helio-astroBERT: We froze the first 10 layers in *helioBERT*, and trained the embedding layer and the last 2 BERT layers on astrophysics texts. Research works in heliophysics and astrophysics share some common glossaries. Instead of collectively training with the texts from these 2 domains, we hierarchically trained the model where it has retained the prior knowledge of heliophysics and will then learn about astrophysics.
Model III. Large helioBERT: Compared to *helioBERT*, a larger amount of heliophysics texts from different sources was used in the training.
Model IV. Large hierarchical helio-astroBERT: We froze the first 10 layers in the *large helioBERT*, and trained the embedding layer and the last 2 BERT layers on a larger amount of astrophysics texts from different sources.

This work is a collaboration between domain scientists and computer scientists. We are able to manually identify meaningful or strong pairs of keywords as a validation of our knowledge graph in the respective scientific domain. As

[2] Impacts from fine-tuning on a Next Sentence Prediction task are left for future studies.

our knowledge graph is fully connected, we sample the results in duplet. We will discuss some of these examples in the result Sect. 5, at times citing the relevant research publications.

4 Experimental Setup

4.1 Data

Our primary data source is NASA's Science Mission Directorate (SMD) dataset, which contains mainly *terms* and *definitions* from 5 scientific domains: Astrophysics, Heliophysics, Planetary, Earth Science, and Biological & Physical Sciences. Examples of such data instances are:

Term: Big Bang theory
Definition: The theory that the Universe 'started' with an event that created time and space, about 13 billion years ago.

Term: Solar Flares
Definition: A great burst of light and radiation due to the release of magnetic energy on the sun. Flares are by far the biggest explosions in the solar system, with energy releases comparable to billions of hydrogen bombs. The radiation from the flare travels at the speed of light, and so reaches Earth within eight minutes. The energy is generally absorbed by Earth's atmosphere, which protects humans on Earth, however, the energy can cause radio blackouts on Earth for minutes or, in the worst cases, hours at a time. The radiation from a flare would also be harmful to astronauts outside of Earth's atmosphere. Some, but by no means all, flares have an associated coronal mass ejection (CME).

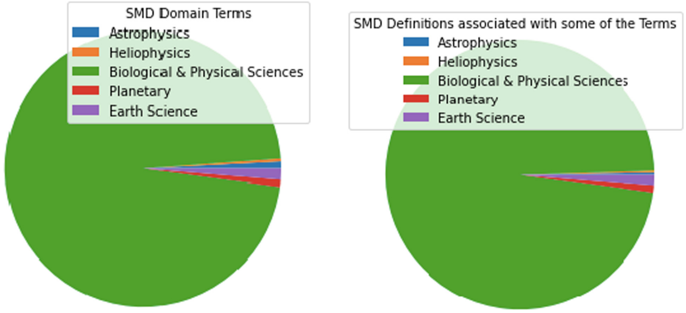

Fig. 3. Pie charts for the number of terms in the SMD dataset per domain (left), and for the number of definitions associated with the terms per domain (right).

Figure 3 shows an overview of our SMD dataset. There are a total of 9,291,463 terms, and 3,096,448 definitions. About 97% of the data come from Biological &

Physical Sciences, while other domains each contribute 0.3–1% of the data. In this work, we will focus on 2 domains: Heliophysics, Astrophysics.

The Unified Astronomy Thesaurus (UAT) data[3] is a table containing 2826 unique terms in the field of astronomy and astrophysics, categorized in 11 levels or hierarchies. For example, if one chooses a level 1 term *Astrophysical processes*, one of the level 2 terms that follows is *Astrophysical magnetism*, then there can be *Magnetic fields* at level 3, and *Primordial magnetic fields* at level 4, etc. The terms become more specific in the higher levels.

Furthermore, we find that there are SMD heliophysics terms which exist in UAT as well: 2% of a total of 29,846 SMD heliophysics terms are in the UAT table. We will refer to these overlapping SMD terms at each UAT level # as "SMD heliophysics level #". Therefore, although these terms are part of SMD heliophysics data, they are less heliophysics-apparent and can be more astrophysics-like. In another word, one can also view this as a shared vocabulary by the two domains.

Basically, the data is used in the following scenarios to: extract relevant arXiv abstract, fine-tune BERT, and extract keywords. The data involved for these purposes are not always the same. In particular, the data we used to fine-tune BERT are (number of texts):
(i) SMD heliophysics definitions (6,336), (ii) SMD astrophysics definitions (9,222), (iii) arXiv abstracts collected using some SMD heliophysics terms (290,316), and (iv) arXiv abstracts collected using some SMD astrophysics terms (200,000).
The data we used for keyword extractions are (number of texts):
(a) arXiv abstracts extracted using SMD heliophysics level 1 (14,227), (b) arXiv abstracts extracted using a particular hierarchy in UAT (5,963), (c) SMD heliophysics definitions (6,336), and (d) SMD astrophysics definitions (9,222).

4.2 Fine-Tuning on BERT: Setup

During the training, we have kept the BERT hyperparameters by default. Following are the specifics for each model training:

helioBERT: Trained on 6,336 SMD heliophysics definitions for 5 epochs.
Hierarchical helio-astroBERT: Trained on 9,222 SMD astrophysics definitions for 5 epochs.
Large helioBERT: Trained on 296,652 texts for 2 epochs. The texts comprise the prior 6,336 SMD heliophysics definitions, and 290,316 arXiv abstracts extracted using a random sample of 100 SMD heliophysics terms.
Large hierarchical helio-astroBERT: Trained on 209,222 texts for 2 epochs. The texts comprise the prior 9,222 SMD astrophysics definitions, and 200,000 arXiv abstracts randomly sampled from a pool of 626,388 arXiv abstracts extracted using a randomly sampled 50 SMD astrophysics terms.

[3] https://astrothesaurus.org, where the list of UAT terms we used are available at https://github.com/astrothesaurus/UAT/blob/master/UAT.csv.

4.3 Keyword Extraction: Setup

For the feasibility of analyzing the results by our domain scientists, for each experiment, we typically select 100 top keywords (tri-grams) extracted by YAKE with the lowest SEQM (i.e. the most relevant ones). As we have considered *a priori* that all the keywords extracted are related, for n keywords selected for further analysis, there will be $\frac{n(n-1)}{2}$ unique pairs of them. We compute the cosine similarities of all the pairs, where a higher cosine value indicates that the pair is more closely connected.

We highlight the following 3 experiments, under two contrasting elements: (i) data sources, and (ii) fine-tuned word embeddings considered.

Experiment I:
The data source is a collection of arXiv abstracts, extracted using a set of terms from a particular hierarchical branch from the UAT table, based on the level 1 term *Astrophysical processes*, level 2 term *Gravitation*, and level 3 term *General Relativity*, and all the terms which follow up to level 11. Hence, there exists a particular knowledge structure here in the data. From this pool of abstracts, we extracted the keywords using YAKE and analyzed the pairs formed out of the top 70 YAKE keywords. In the next section, we will show the comparison of the connections resulted using the embeddings from *hierarchical helio-astroBERT* against its *large* version.

Experiment II:
This experiment plans to show how BERT which has learned some heliophysics handles the more astrophysical data or the shared vocabularies between heliophysics and astrophysics domains. The data source is a collection of arXiv abstracts, extracted using a set of terms which we refer to as "SMD heliophysics level 1". This experiment compares the resulted graphs of scientific pairs using the word embeddings from *helioBERT* and *large helioBERT*. Top 89 YAKE keywords were selected.

Experiment III:
The data source is simply a collection of SMD heliophysics definitions and SMD astrophysics definitions. This is to examine the connectivity between the two scientific domains. Top 172 YAKE keywords were selected in this experiment.

We summarize the background details of the experiments in the following Table 1:

Table 1. Characteristics of the experiments. Shorthand for the model names: *lhhaB*: *large hierarchical helio-astroBERT*, *hhaB*: *hierarchical helio-astroBERT*, *lhb*: *large helioBERT*, *hb*: *helioBERT*, where their word embeddings are used.

Expt.	# unique keywords	# pairs	SEQM	Embedding	# pairs with cosine sim., α
I	70	2415	$(6.4\text{–}25)\times10^{-5}$	*hhaB*	203 ($\alpha > 0.6$)
I	70	2415	$(6.4\text{–}25)\times10^{-5}$	*lhhaB*	334 ($\alpha > 0.6$)
II	89	3916	$(1.0\text{–}9.9)\times10^{-4}$	*hB*	20 ($\alpha > 0.8$)
II	89	3916	$(1.0\text{–}9.9)\times10^{-4}$	*lhB*	41 ($\alpha > 0.8$)
III	172	14,706	$(1.0\text{–}9.9)\times10^{-4}$	*hhaB*	5751 ($\alpha > 0.5$)

5 Results

Our fully-connected knowledge graphs are massive even with under 200 keywords/nodes. As there is no ground truth to verify all the connections, it is useful that we could single out a number of interesting or true example pairs for discussions. We present these examples here and furthermore, we provide some relevant references accordingly (externally linked).

Result from Experiment I:
We compare the representations resulted from two word embeddings: *hierarchical helio-astroBERT* and its *large* version, focusing on the pairs whose cosine similarities α are higher than 0.6. By the *hierarchical helio-astroBERT* embeddings, we point out in particular in Table 2 some interesting example pairs.

Table 2. Example pairs highlighted from Experiment I (with *hhab* embeddings).

(*James Webb Space*, *Schwarzschild black hole*): $\alpha = 0.7112$
(*Generalized Uncertainty Principle*, *Einstein General Relativity*): $\alpha = 0.601$ (reference)

Although black hole of precisely Schwarzschild is rather too specific (theoretical), black holes can be connected with James Webb, as data from Webb can be used to study e.g. the growth rate of supermassive black holes (reference). Also, we find that the three physics journals (*Phys. Rev. Lett, Proc. Roy. Soc, Phys. Dark Univ.*) are connected to each other with a cosine similarity of more than 0.7. Table 3 shows a list of results with the highest cosine similarity.

Table 3. Top results from Experiment I (with *hhab* embeddings).

(*Phys. Rev. Lett.*, *Laser Interferometer Gravitational-wave*): $\alpha = 0.8637$
(*polynomial curvature invariants*, *Gravitational Lensing Experiment*): $\alpha = 0.8703$
(*Small Magellanic Cloud*, *Large Magellanic Cloud*): $\alpha = 0.8815$
(*Counterpart All-sky Monitor*, *Laser Interferometer Gravitational-wave*): $\alpha = 0.9624$
(*Cosmic Microwave Background*, *Phys. Rev. Lett*): $\alpha = 0.9812$

While, by the *large hierarchical helio-astroBERT* embeddings, we point out in particular in Table 4:

Table 4. Example pairs highlighted from Experiment I (with *lhhab* embeddings).

(*James Webb Space*, *Cold Dark Matter*): $\alpha = 0.8701$
(*Cold Dark Matter*, *Webb Space Telescope*): $\alpha = 0.6076$
(*Event Horizon Telescope*, *Massive Black Hole*): $\alpha = 0.6124$

These are again convincing pairs. The data from James Webb will help to verify the existence of cold dark matter (reference). Note the changes in the cosine similarity α for the pair containing *Cold Dark Matter* when its partner is *James Webb Space* or *Webb Space Telescope*. Even though the keyword extraction by YAKE is not complete as in *James Webb Space Telescope* (as we had required for tri-gram), the associated cosine similarity is still high. Table 5 shows a list of results with the highest cosine similarity.

Table 5. Top results from Experiment I (with *lhhab* embeddings).

(*Fourth Mexican School*, *Laser Interferometer Gravitational-wave*): $\alpha = 0.8828$
(*Gravitational Lens Astrophysics*, *Einstein General Relativity*): $\alpha = 0.8864$
(*Massive Black Hole*, *Extremely Compact Objects*): $\alpha = 0.9388$
(*Expansive Nondecelerative Universe*, *Field Dark Matter*): $\alpha = 0.9489$
(*Interferometer Space Antenna*, *Cosmological Gravitational Lensing*): $\alpha = 0.9767$

By narrowing down our scope in the text corpus to gravity in general (built off a particular hierarchy in UAT), we are able to observe extremely informative pairs right from this research area.

Result from Experiment II:
We compare the representations resulted from two word embeddings: *helioBERT* and its *large* version, focusing on the pairs whose cosine similarities α are higher than 0.8. By the *helioBERT* embeddings, we point out in particular in Table 6 the example pairs found. Table 7 shows a list of results with the highest cosine similarity.

Table 6. Example pairs highlighted from Experiment II (with *hb* embeddings).

(*Spitzer IRAC based*, *Big Bang theory*): $\alpha = 1$ (reference)
(*phantom divide line*, *quantum gravity community*): $\alpha = 0.8373$ (reference)

Table 7. Top results from Experiment II (with *hb* embeddings).

(*neutral massive fields*, *phantom divide line*): $\alpha = 0.9494$
(*main modern developments*, *CMB anisotropy data*): $\alpha = 0.9609$
(*GOYA Survey imaging*, *QSO absorption line*): $\alpha = 0.9892$
(*understanding current theories*, *Long Baseline Array*): $\alpha = 1$
(*spinning fluid embedded*, *Supernova Legacy Survey*): $\alpha = 1$

While, by the *large helioBERT* embeddings, we point out in particular in Table 8 some example pairs.

Table 8. Example pairs highlighted from Experiment II (with *lhb* embeddings).

(*asymmetric dark matter*, *Standard Model imposed*): $\alpha = 1$
(*Hubble Volume N-body*, *Long Baseline Array*): $\alpha = 1$ (reference)
(*phantom divide line*, *main modern developments*): $\alpha = 0.9455$
(*phantom divide line*, *dark matter halo*): $\alpha = 0.8635$
(*main modern developments*, *dark matter halo*): $\alpha = 0.8633$

Although the keywords *asymmetric dark matter* and *Standard Model imposed* are paired with cosine similarity 1 (Table 8), it should not be taken literally, as we need to go beyond the *Standard Model* in order to explain dark matter. Table 9 shows a list of results with the highest cosine similarity.

Table 9. Top results from Experiment II (with *lhb* embeddings).

(*observed Velocity Dispersion*, *X-ray analyses lead*): $\alpha = 0.9513$
(*Cosmological General Relativity*, *cosmic microwave background*): $\alpha = 0.9891$
(*Hartle-Hawking No-Boundary Proposal*, *Density linear perturbations*): $\alpha = 1$

Interestingly, these highlighted examples show that BERT with only heliophysics knowledge is able to identify with a good indication of the strength of the relations on astrophysical contents such as *CMB*, *Big Bang theory*, to name a few.

Result from Experiment III:
By *hierarchical helio-astroBERT* embeddings, we find elements from the 2 domains connected with more than 0.5 cosine similarity, in particular in Table 10 we point out some example pairs.

Table 10. Example pairs highlighted from Experiment III (with *hhab* embeddings).

(*Martian satellite Phobos*, *Heliospheric Solar Magnetospheric*): $\alpha = 0.5349$
(*Measurements CME motion*, *Heliospheric Solar Magnetospheric*): $\alpha = 0.5284$ (reference)

There are indeed studies on interactions between solar wind and the Mars-Phobos (reference). Table 11 shows a list of results with the highest cosine similarity.

The cross-domain relations that we find are encouraging. The terms in SMD dataset are usually more technical, very specific to a smaller research community, as it involves for example names of instruments. Hence the connections established here are more technical than conceptual.

Table 11. Top results from Experiment III (with *hhab* embeddings), with $\alpha = 1$.

(*South Pacific Ocean*, *Synthetic Aperture Radar*) (reference)
(*Geocentric Equatorial Inertial*, *Lunar Reconnaissance Orbiter*)
(*SSA Space Weather*, *Explorer Mission satellite*)
(*Sciences Laboratory Facility*, *Polar Cap Indices*)
(*Naval Observatory Astronomical*, *Small Explorer Project*)

6 Discussions

In our approach, firstly, the strength of the cosine similarity is rather relative to the scope, and the size of the text corpus being considered during both the fine-tuning stage and keyword extraction. The scope of the corpus can be inferred from the set of terms we used to extract the relevant texts. It is non-trivial to determine exactly the relatedness of the entities, as the rank could change according to the depth and width of the respective research area, or its collection of research papers. On the other hand, one can see from Table 1 that a larger fine-tuned language model tends to produce a larger number of pairs at the same level of cosine similarity.

Secondly, there is not a clear best language model among those we proposed. The reason is related to the first point. Here we have looked at the aspect of hierarchical training, and also the results from using a different text size in fine-tuning. We do find interesting outputs from all the cases considered. Most importantly, the type of texts where the keyword extraction is performed plays a crucial role in producing some of the strongest relations: there is an implicit term hierarchy in the texts from Experiment I; shared scientific terms between the 2 domains from Experiment II; purely a combination of the technical terms from the 2 domains from Experiment III. Thus, we think that the models could as well complement each other in completing a knowledge graph. For more discussions about the related challenges, see e.g. [13].

7 Conclusions

We propose an embedding-based method to construct a knowledge graph in heliophysics and astrophysics domains, utilizing the cosine similarities computed from the word embeddings of the respective domain-specific BERT. Bypassing the need for a fine-tuned named entity extraction or such labeled training dataset, and out of a pool of texts selected based on a set of controlled scientific terms, our constructed knowledge graph is able to present many convincing relations of scientific concepts or terms in and across the domains. Moreover, our fine-tuned BERT models can also be used for other downstream tasks such as NER.

For future work, we plan to improve our validation method by using automatic metrics proposed in such as [14,15] which are based on declarative and mined rules, in addition to human-in-the-loop. We also plan to extend our method to study the synergies with the remaining scientific domains of planetary, earth science and biological & physical sciences.

In our fine-tuning of BERT and also in the arXiv abstract extraction (arXiv content to enrich our SMD dataset), scientific terms from the SMD dataset have been actively involved. Our ultimate goal is to develop our approach into a useful search tool for domain scientists to assist them in their research, and for integration into NASA's SMD data system.

Acknowledgements. This work has been enabled by the Frontier Development Lab (FDL) (https://frontierdevelopmentlab.org). FDL is a collaboration between SETI Institute and Trillium Technologies Inc., in partnership with NASA, Google Cloud, Intel, NVIDIA, and many other public and private partners. This material is based upon work supported by NASA under award No(s) NNX14AT27A. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration. The authors would like to thank the FDL organizers, the SETI institute, Trillium, the FDL partners and sponsors, and the reviewers for their constructive comments during the research sprint. The authors thank Sergi Blanco-Cuaresma and Felix Grezes for helpful discussions. Finally, the authors also thank Google for providing access to computational resources without which this project would not have been possible.

References

1. Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. Semant. Web **8**(3), 489–508 (2017)
2. Gao, J., Li, X., Xu, Y.E., et al.: Efficient knowledge graph accuracy evaluation. VLDB Endow. **12**, 1679–1691 (2019)
3. Mintz, M., Bills, S., Snow, R., et al.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011 (2009)
4. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
5. Devlin, J., Chang, M.-W., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)
6. de Jongh, M., Druzdzel, M.J.: A comparison of structural distance measures for causal Bayesian network models. In: Recent Advances in Intelligent Information Systems, Challenging Problems of Science, Computer Science Series, pp. 443–456 (2009)
7. McGranaghan, R., Klein, S.J., Cameron, A.: The NASA Heliophysics KNOWledge Network (Helio-KNOW) project (an essay for the Space Data Knowledge Commons)
8. Bentley, R., Brooke, J., Csillaghy, A., et al.: HELIO: discovery and analysis of data in heliophysics. Futur. Gener. Comput. Syst. **29**, 2157–2168 (2013)

9. Efthymiou, V.: CosmOntology: creating an ontology of the cosmos. In: DL4KG: Deep Learning for Knowledge Graphs Workshop (ISWC 2022) (2022)
10. Grezes, F., et al.: Building astroBERT, a language model for astronomy & astrophysics. arXiv preprint arXiv:2112.00590 (2021)
11. Beltagy, I., Lo, K., Cohan, A.: SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620 (2019)
12. Campos, R., Mangaravite, V., Pasquali, A., et al.: YAKE! keyword extraction from single documents using multiple local features. Inf. Sci. **509**, 257–289 (2020)
13. Timmer, R.C., Mark, M., Khoo, F.S., et al.: NASA science mission directorate knowledge graph discovery. In: WWW 2023 Companion: Companion Proceedings of the ACM Web Conference 2023, pp. 795–799 (2023)
14. Ortona, S., Meduri, V.V., Papotti, P.: RuDiK: rule discovery in knowledge bases. In: Proceedings of the VLDB Endowment, pp. 1946–1949 (2018)
15. Tanon, T., Bourgaux, C., Suchanek, F.: Learning how to correct a knowledge base from the edit history. In: Proceedings of WWW 2019: The World Wide Web Conference, pp. 1465–1475 (2019)

Author Queries

Chapter 15

Query Refs.	**Details Required**	**Author's response**
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Please note that the footnote has been set in the following sentence "This work has been enabled...", as footnotes are not allowed in Acknowledgements.	