Response to the Reviews and Decision

Title: Towards Semantically Enriched Embeddings for Knowledge Graph Completion

Manuscript Tracking Number:
673-1653

Authors:
Mehwish Alam
Frank Van Harmelen
Maribel Acosta

Date: 23.02.2024

We would like to thank the reviewers for their constructive comments. In the following, we are addressing the concerns of each of the comments and also modifying the manuscript when needed. All the changes are highlighted in blue in the new version of the manuscript.

Response To Reviewer #1

Reviewer Comment #1

> *1. In general, if this wants to be a survey paper, I believe that the knowledge graph embedding summary could benefit from some additional details on what models actually do and a more general introduction to embeddings (e.g., how is a scoring function used to generate a prediction).*

Response

We have reformulated the paper type as a position paper, where we summarize the advances in KG embeddings for knowledge graph completion, and providing critical insights into limitations and gaps of existing approaches in exploiting different aspects of KGs (i.e., ontological information or external sources). We have clarified this in the abstract and introduction of the manuscript. We also followed the reviewer's suggestion, and added at the beginning of Section 3.1 a paragraph about the details mentioned by the reviewer with a reference to the survey that describes the relevant algorithms. We have also changed the order of Sections 3.2 and 3.3. Section 3.2 is now Section 3.3 and vice versa.

Reviewer Comment #2

> *2. There are some papers that have not been cited (e.g., `https://arxiv.org/pdf/1903.05485.pdf` on multi-modal kg embeddings). I also understand that this might not be the entire focus of the paper (as some of this info can be found in other survey papers), but it also depends on how much this paper wants to be a survey paper.*

Response

Thank you for your comment. However, we think that there is a huge literature on these methods out there. Since the idea of this article is to give a teaser to these methods, we have referred to a detailed survey on these methods, please refer to the paper [4] in the bibliography.

Reviewer Comment #3

> *2. If it is not possible to extend the survey due to page constraints, I'd still try to reorganize it and provide more details on the vision and the discussion part (4.3 is one of the few sections with a section-specific discussion). For example, I find the recommendations section useful, but I would again extend it and present the content in a more structured format (e.g., a figure or a table). This kind of format could help in*

conveying the main ideas quickly.

Response

Thank you for your comment. We have extended the recommendation section (now Section 7.2). The evaluation section is itself a summary as well as a discussion.

Reviewer Comment #4

2. I find the paper a bit terse, as some things are briefly mentioned but not really explained. There are many instances of this problem. "GenKGC [61] converts the KG completion task to a sequence-to-sequence (Seq2Seq) generation task. The incontext learning paradigm of GPT-3 learns correct output answers by concatenating the selected samples relevant to the input." - what are the selected samples here? In addition to this, the next sentence starts with "GenKGC similarly" but I am not sure what "similarly" refers to.

Response

Thank you for pointing it out. We have modified the description in Section 4 (previously Section 4.1) and everywhere else whenever possible.

Reviewer Comment #5

2. Section 5 is very short and some paragraphs would require better structure and better organization.

Response

Section 5 is now organized into three subparts, i.e., "Benchmark Datasets", "Metrics" and "Evaluation and Reported Results". We have also added a few more details on each of these aspects.

Reviewer Comment #6

2. "ELEm has been evaluated on the Protein-Protein Interaction (PPI) dataset for the LP task. However, the successor algorithms introduced more appropriate datasets such as Gene Ontology (GO)" - why are these more appropriate and why does this make a difference?

Response

We have added a sentence targeting this comment in the Evaluation section (Section 6).

Reviewer Comment #7

> *2. Then, in the same section, the paper describes the results in Ruffinelli et al., which are important results in the KG embeddings evaluation, but I am not sure if this is the best way or section to introduce them. Also, since there have been some attempts at providing more uniform benchmarking utilities such as* `https://github.com/pykeen/pykeen`*, I think it might be worth mentioning these (the authors briefly introduce the paper in ref [80] but since this is an evaluation setup section I think more details could help).*

Response

(Now) Section 6 on "Existing Evaluation Settings" has been restructured and the reference to Pykeen has been added.

Reviewer Comment #8

> *2. It is sometimes unclear to me if the focus is on Knowledge Graph embeddings or LLMs. It seems to me that LLMs appear mostly in one section, but from the introduction, I'd have expected a larger analysis and more details in the discussion section. From the title the focus should be semantically enriched embeddings, but the focus seems on more general knowledge graph embeddings to me.*

Response

We agree with the point since also the paper states that LLMs are a good source of background knowledge. Following that, the big chunk of text on LLMs has been removed from the introductory section to focus more on the problem at hand.

Reviewer Comment #9

> *2. Additional ref and dataset for inductive link prediction:* `https://arxiv.org/abs/2203.01520` *On multi-modality:* `https://arxiv.org/pdf/1903.05485.pdf`

Response

ILPC2022 (`https://arxiv.org/abs/2203.01520`) has been added to Table 3. For `https://arxiv.org/pdf/1903.05485.pdf`, in principle, we could add this citation but we have referred to the surveys which are more complete than our brief section on multimodal knowledge graph embeddings and we have mentioned only a few selected works. If we need to add all the algorithms that are on this problem this will be yet another survey on knowledge graph embeddings which is not the main focus. These categories of algorithms are introduced to give an overall view with references to surveys if existing. We hope it clears up the issue.

Reviewer Comment #10

> *2. RoBerta should be RoBERTa*

Response

This instance has been corrected.

Reviewer Comment #11

> *2. Table 3. I think this is a useful table, as it summarizes many of the different evaluations. Maybe it would be good to add the references for the datasets?*

Response

The references to the datasets have been added to the paper.

Response To Reviewer #2

Reviewer Comment #1

> *I believe the style of the abstract could be changed to match the tone of the rest of the paper. The bottom half reads a bit like an enumeration of sections, and it would be more valuable if it actually read like a summary of important conclusions, which is the great contribution of the paper.*

Response

Thanks for the suggestion. The abstract has been modified accordingly.

Reviewer Comment #2

> *I think the Introduction takes too long to reach the crux: knowledge graph completion. The reader must go through a paragraph and a half about LLMs before hitting the KG completion. I think it would benefit from a complete restructuring to place more emphasis on (1) why KG completion is an important task and (2) why are semantically enriched KG embeddings important to support it. The paragraphs about LLMs feel out of place. LLMs are ONE way to introduce more richness, are perhaps THE way to go in the future, but the problem formulation needs to be well-established before going into those particulars.*

Response

We have removed the paragraph on LLMs from the introduction.

Reviewer Comment #3

> *Section 3 is titled Knowledge Graph Embedding Algorithms but in sections 3.2 and 3.3 other types of methods are presented. A clear restructuring would be needed to cover both the related work on KG embeddings and that on KG completion.*

Response

The title of the section has been changed to "Knowledge Graph Completion using Embeddings".

Reviewer Comment #4

> *I miss a table in section 3 organizing the work in KG completion into the types of embeddings used, and use of contextual/external information.*

Response

Due to space constraints, we cannot include such a table in the manuscript. We hope that the new structure of the paper makes it easier to follow the types of embeddings that are discussed.

Reviewer Comment #5

> *Comment on "The algorithms discussed so far consider only statements in the ABox for generating KG embeddings and performing LP." In Table 2 you present many approaches that embed the TBox. They could in principle be applied to KGs, including the ABox. Some more discussion on this point would be nice.*

Response

Thank you for your comment. In this line by the phrase "so far" we were referring to the algorithms discussed in Section 3. The phrase has been edited in the manuscript.

Reviewer Comment #6

> *Comment on: "A vast amount of knowledge is captured by LLMs, type hierarchy and the expressivity of the description logic axioms has not been considered. The subsequent sections focus on these aspects of LP." This sentence feels out of place. It would require a lot more context to make sense. For one, there is no shortage of embedding methods that consider type hierarchies (all random-walk methods do, ontology-oriented methods do, and there is nothing to prevent other KGE methods to also explore the TBox, even if they interpret those triples just as ABox triples). I would phrase the LLMs as an opportunity and the others as limitations. While I agree that ignoring DL*

axioms can be seen as a limitation of KG embedding methods, I argue that not exploring LLMs isn't. LLMs are "external sources of knowledge" which is a fundamentally different aspect.

Response

The sentence has been rephrased to convey its meaning correctly. We agree with the reviewer that ignoring DL axioms is a limitation: KGE that interpret TBox triples as ABox triples do not achieve high performance in reasoning tasks such as membership and subsumption prediction (see results reported by Chen et al. [78]). We have re-organized the paper to better present this. The article progressively discusses which kinds of semantics/external knowledge are taken into account for link prediction and near the end it concludes what is missing based on these works that are being discussed throughout the article.

The paragraph on using LLMs for KG completion has been placed as a separate section.

Reviewer Comment #7

In 4.1 I miss a clear value proposition on how LLMs can capture semantics. The section is a survey of existing works, but it lacks a clear way to place the mentioned works in the context of capturing semantics. There is also no discussio on this, since the discussion portion of this section only covers 4.2 and 4.3.

Response

It has been introduced as a separate section (Section 4) since it serves more as a technique and background knowledge as compared to the ontological axioms which is the focus of this study.

Reviewer Comment #8

One of the issues that is raised is the size of benchmarks. It would be helpful if Table 3 also included the size of the datasets.

Response

Since different algorithms consider different kinds of axioms, the statistics reported for these algorithms consider different axioms which makes it very difficult to report unified statistics. However, the comment is very useful and we plan on taking this into account for future work originating from this article.

Reviewer Comment #9

Table 3 is missing citations to the benchmarks and ontologies it refers to.

Response

The citations have been added.

Reviewer Comment #10

> *Sem@K, although a step in the right direction, does not address the problem of genuinely new links. Some insight on how this could be tackled would be nice.*

Response

This comment has been partially discussed in the following chunk of the last section:

"*sem@k* aims to rectify this to some extent, although it remains unclear at the moment how much of this evaluation can be done without expensive human annotation of gold standards."

This line indicates that human evaluation for new links is needed, however, it is expensive to obtain this kind of ground truth. This line provides an insight into the question posed in this review point.

Reviewer Comment #11

> *Missing citations to relevant works in these areas: "Other KGs can also provide such background knowledge, leading to an interesting blurring between the tasks of KG linking and KG completion. Recent work on exploiting the temporal evaluation of a KG as the source of information is another example of using information outside the KG for KG completion."*

Response

The citation on integrating multiple knowledge graph embeddings has been added which is in synchronization with the point made in the quoted line [1].

[1] M. Baumgartner, D. Dell'Aglio, H. Paulheim, and A. Bernstein, Towards the Web of Embeddings: Integrating multiple knowledge graph embedding spaces with FedCoder, Journal of Web Semantics (2023).

Response To Reviewer #3

Reviewer Comment #1

> *Regarding the related work part in Introduction, it looks quite short, with some recent important survey, position and vision papers missing. One example is:* `https://arxiv.org/abs/2308.06374`*, which introduces the combination of LLMs and knowledge graph. Although it does not exactly match the purpose of this paper, it covers the part of knowledge graph completion using LLMs. Briefly, I think more should be discussed on the recent survey, position and vision papers related to knowledge graph embedding/completion, and the difference of this paper in comparison with the existing ones.*

Response

Thanks for the suggestions. We have included `https://arxiv.org/abs/2308.06374` and `https://arxiv.org/pdf/2306.08302` in the related work, which discuss complementary aspects the interplay of KGs and LLMs.

Reviewer Comment #2

> *The title of Section 3 is "Knowledge Graph Embedding Algorithms", but the subsections are organised according to the knowledge graph completion task. Why not directly name Section 3 as "Knowledge Graph Completion Tasks via Embeddings" or some title similar.*

Response

Thank you for the comment. The title of the subsection has been changed (Section 3).

Reviewer Comment #3

> *The title of Section 4 is "Towards Capturing Semantics in Knowledge Graph Embeddings". This means the subsection will introduce the works from the perspective of "Semantics" utilised. But it seems Section 4.1 is to introduce the methods of using the LLM technique. I can understand most of the LLM-based method for knowledge graph completion utilise the semantics of literals, but that may not be 100% correct. I would suggest to describe methods in Section 4.1 from the perspective of using literal semantics (LLM is not a kind of knowledge graph semantics, but a technique). Section 4 .2 and 4.3 describe the works from the perspective of semantics, but for Section 4.3, why not directly name the semantics e.g., TBox in the title, instead of using "Semantically Rich Embeddings" ("rich" can refer to many kinds of semantics).*

Response

Thank you very much for the comment. The Large Language Models (generative models) such as ChatGPT when prompted can provide the predictions for head, tail, and relation. In that sense, they are not focused on literals. However, we agree with the comment that they serve as a technique and background knowledge. To meet this critique we have now introduced Large Language Models for Knowledge Graph Completion as a separate section (Section 4) and previously Section 4 on "*Towards Capturing Semantics in Knowledge Graph Embeddings*" is now Section 5. To reflect the aspect of TBox when referring to semantics, we have renamed the new Section 5.2. to "O*ntology-Enriched Knowledge Graph Embeddings*".

Reviewer Comment #4

> *Section 5 introduces evaluation setting, but it is not complete. I would suggest to list and introduce the commonly used benchmarks for knowledge graph embedding benchmarking, such as Wikidata5M and FB15K.*

Response

These have been listed and referenced in Table 3 along with the references. A reference to the paper comparing the LP datasets has been added. Due to page limitations, it does not seem feasible to add such a detailed description since the list is very long.

Reviewer Comment #5

> *As a position, more should be discussed on the future directions. Section 6.2 gives some future direction discussion, but it seems to be quite abstract, being short of concrete technical challenges. For example, the future direction sentence from 46-47 mentions "challenges of inductive setting". But what kinds of challenges? What challenges have been partially investigated and what haven't?*

Response

The challenges already targeted in the inductive setting have already been discussed in the dedicated section (Section 3.2) and the other challenges are discussed in Section 3 (Discussion) and finally referred to in the last section *"Bias against external knowledge"*.

Reviewer Comment #6

> *I understand this domain has so many papers to cite. As a position paper, it is not necessary to cite all of them. But some important and representative ones should be considered. For example, for inductive knowledge graph completion, unseen entities, unseen relations, and both unseen entities and unseen relations should be considered. Here are two papers for consideration on this topic (and the categorisation):* `https://dl.acm.org/doi/abs/10.1145/3442381.3450042` `https://arxiv.org/pdf/2210.03994.pdf`. *Similarly, the other topics could also have method categorisation, and have representative papers cited for each category.*

Response

Thank you very much for your comment. We have added the paper `https://arxiv.org/pdf/2210.03994.pdf` in Section 3.3 and `https://dl.acm.org/doi/abs/10.1145/3442381.3450042` in Section 5.2 (previously 4.2).

Reviewer Comment #7

> *The category column of Table 3 is quite strange. They are not categorised from on dimension. The current categories may have much overlap, e.g., LLM-based methods and inductive link prediction.*

Response

Thank you for your comment. In Table 3, the categories are defined based on the categorization given in the rest of the paper and the tasks on which these categories of the algorithms are evaluated. For the sake of understandability we have further added fine-grained task divisions along with the datasets used for those tasks. The explanation of the table has also been modified.
