View Reviews

Paper ID
238

Paper Title
Realistic Re-evaluation of Knowledge Graph Completion Methods [Experiments and Analysis]

Track Name
Research Paper First-Round

Reviewer #1

Questions

1. Overall evaluation
Weak Reject

2. Reviewer's confidence
Knowledgeable

3. Novelty
Medium

4. Importance: select all that apply:
The paper contains controversial ideas and/or will generate interesting discussion

5. Summary of contribution (in a few sentences)
The paper presents experimental studies on using knowledge-graph embeddings
for so-called KG completion, that is, predicting missing but valid KG triples.
The paper makes a few very good points about missing realism in the tasks,
shortcomings of state-of-the-art methods and pitfalls and misinterpretations
of benchmark results.
I believe these are important messages to make.
The experiments that substantiate these points are, as far as I can tell,
carried out thoroughly.

6. List 3 or more strong points, labeled S1, S2, S3, etc.
S1 interesting experimental observations, potentially of "eye-opening" but also controversial nature

7. List 3 or more weak points, labeled W1, W2, W3, etc.
W1 writing style neds to be toned down and made more concise
W2 scope of benchmarks and experiments needs to be expanded

8. Detailed evaluation. Number the paragraphs (D1, D2, D3, etc.)
So this seems to be a candidate for an insightful experimental paper,
which would be good to have at the conference.
However, the paper has a number of significant limitations and flaws, in its current form.
So it does need substantial revision before it can be considered for publication.

1) Writing style:
The paper is written in a very aggressive style, making the same
arguments over and over again and putting them in strong wording.
The intro runs almost up to the end of page 3: too long and repetitive.
Often, sentences are underlined, in addition to strong wording.

The phrase "This study can have large potential impacts" is in
boldface: the impact of a paper comes from its readers and the community,
not from the authors saying it strongly.
Another example of overly aggressive wording is "This study puts into question
the foundational principles of the field of ..." - a very strong statement
that borders on discrediting an entire sub-community.
I suggest that the writing style undergo substantial revision,
to tone down the language and make points more concisely in a purely technical way.

2) Scope of benchmarks and experiments:
A large part of the paper is about the shortcomings of the FB15k benchmark.
It boils down to discussing FB15k vs. the newer FB15k-237.
But there are quite a few alternative datasets for experimental evaluation
of KG-completion methods. See, for example, the various data used in the ConvE
paper (Dettmers et al. 2018) or in the recent paper by Lin, Socher and Xiong
in EMNLP 2018.
What about these additional benchmarks? What about looking at a variety of
them jointly - would this bring out a different, more refined and informative picture?
The paper makes it sound as if almost all of the KG-completion community
has focused on FB15k and only FB15k. But hasn't the community already
abandoned the older FB15k to a large degree?
This entire point of choosing benchmarks
needs to be discussed in a broader and clearer way.
Experimental comparison should consider at least one other choice of dataset
beyond the FB15k variations.

3) Spectrum of idiosyncracies:
The paper repeatedly discusses the pitfalls from focusing on
a small number and types of predicates like inverse relations,
Cartesian product relations and other redundant relations.
These points are very good, but they could be made more concisely.
In addition, the paper should discuss further issues as well, including
predicting triples via paths of predicates (multi-hop cases).
Also, the point made about higher-arity predicates (CVT nodes in Freebase,
Wikidata has a similar construct) is very interesting;
so it would be great to cover such cases in experiments as well
(to the degree that is possible - but some recent works address
some of these cases).

4) Overriding usefulness of KG-completion:
The paper makes a good point about the missing realism of
KG-completion as a technical task. I fully agree with this, and
would even make it stronger: KG-completion is meant to help
a KG curator, but predictions with the kind of reciprocal ranks that these
methods achieve are hardly helpful. No human curator would want
to sift through dozens of spurious predictions before coming upon
one useful.
Perhaps, the paper could add a discussion on these human-user aspects
as well (where users are curators).

5) Bibliography (minor point):

The authors should carefully revisit the 2018 and 2019 publications
on this topic, to ensure that there is indeed good coverage of recent works.
I noticed, for example, that the EMNLP 2018 paper by Lin, Socher and Xiong
is missing. Also, there is a TKDE 2017 survey paper by Wang et al.,
which should probably be cited. There is likely more.
Of course, nobody expects a perfect bibliography, but with the strong
critique of prior works that the paper is making, the bibliography
better be reasonably foolproof with regard to 2018/2019 papers.

9. Candidate for a revision?
Yes

10. Required changes for a revision, if appropriate (labeled R1, R2, R3, etc.).
address points listed under weaknesses

12. Comments on the revised paper (if appropriate)
The authors have done a good job on revising the paper,
especially regarding the extension of its experimental studies.
By and large, I am satisfied with the paper now, and
recommend that it be accepted.

There are some points, though, that the final version
should incorporate towards a crisper presentation:

1) The intro has become too long now, because the revision
only added text, and the conclusion has been untouched.
The intro would gain by being trimmed to bring out
the gist of the paper more clearly.
The conclusion should be reconsidered to give the take-home
message more clearly.
To me, the take-home from the paper is twofold:
A) most importantly, the existing benchmarks are too easy
and thus insufficient - better stress-test benchmarks are needed
B) once the empiricial results are re-assessed when considering
point A, it becomes clear that the state-of-the-art on KG completion
via embedding-based link predictions is not good and far
from being practically viable.
These two messages, especially message A, must come across more clearly.
Everything else is of secondary nature, and the paper should avoid
being cluttered with too many secondary issues (especially in the intro).

2) The intro lists some points A1, A2, A3 (called "alarming ideosyncrasies")
and results R1 to R3.
Both of these lists should be itemized, so as to stand out from the text.
What is the difference between the A list and the R list?
For example, A2 is pretty much about the experimental results - so isn't
this an R item?
The order in the text is also confusing: A1, A2, R1, R2, R3 and then A3.
Clarify and improve!
Also, keep the wording scientific, avoid unnecessary words such as
"alarming".

3) As for prior work (mentioned in the intro), the paper gives credit to references 1 and 35. I believe that reference 9 (Dettmers et al.) deserves similar credit. It already goes a long way on revealing these ideosyncracies in the prevalent benchmarks.

4) As for evaluation metrics, why do you include both MR and MRR and both FMR and FMRR? Wouldn't MRR and FMRR be sufficient? Showing less in the tables and fewer numbers in the text would make the paper look nicer and easier to read, without losing any message.

5) It would be good to clearly identify which of the benchmark datasets is least prone to the discussed ideosynracies (in the test fold), that is, avoiding bijective properties, Cartesian-product cases etc.
In other words, which of the benchmarks comes closest to the desired stress test for KG completion?
Wouldn't the Yago3-10 data slice (emphasize that this is a small fraction of the actual KG) introduced in reference 9 (Dettmers et al.) be a reasonably good case? Affiliations, musical roles, locations of organizations etc. are many-to-many relations, aren't they?
Please discuss this a bit more. Right now, the text alludes to big redundancy in the Yago3-10 data (affiliations and playsFor, see page 14), but is this really the case? Don't these two relations differ in their range types (playsFor only for athletes, affiliations for all other person types, especially scientists)?
Even if this recent benchmark does indeed suffer from the same flaws as the older ones, please discuss what can be done about it (e.g., remove a few relations from that data, or add more, or derive a benchmark from Wikidata or ...). Be constructive!

13. Recommended decision for the revised paper (if appropriate)
Accept

Reviewer #2

Questions

1. Overall evaluation
Weak Reject

2. Reviewer's confidence
Expert

3. Novelty
Medium

4. Importance: select all that apply:
SIGMOD attendees will learn something interesting from the paper
The paper contains controversial ideas and/or will generate interesting discussion
The paper is likely to influence other research in the community

5. Summary of contribution (in a few sentences)

This paper investigates, with great detail, the effect of hidden biases in some of the widely used benchmarks for link prediction tasks.

The experiments reveal a previously unknown type of relations that lead to information leak between training and test splits. They also experimentally show some of the problems of evaluating learning methods in KGs that have the open-world assumption.

6. List 3 or more strong points, labeled S1, S2, S3, etc.

S1. The paper is very well written and organized in a nice and easy to follow manner.

S2. This paper sheds light on, probably one of the most important reasons, why some embedding based link prediction models perform better than others on one of the widely used benchmark datasets. It also makes available some previously unknown information-leakages between train/test splits of these benchmarks.

S3. It compares rule-based models and embedding based models with more details and analysis.

S4. It provides code and reproducibility.

7. List 3 or more weak points, labeled W1, W2, W3, etc.

W1. This paper is highly overlapping with "Re-evaluating Embedding-Based Knowledge Graph Completion Methods. Farahnaz Akrami et al, Conference on Information and Knowledge Management 2018". To the extent that many of the results and tables are directly borrowed from there.

W2. Authors make claims without backing them with references and at times false claims/deductions.

W3. Lack of experiments/results from the state of art embedding methods and datasets.

8. Detailed evaluation. Number the paragraphs (D1, D2, D3, etc.)

D1. (W1 ctd.) I suggest referring the reader to the 2018 paper, reduce the overlap and include some of the more recent embedding models (e.g. rotatE 2019, TuckER 2019, HyperKG 2019). Some of this work (e.g. HyperKG) don't even continue using FB15K and only use FB15K-237, it would be useful to reproduce their results on FB15K and analyze them. It would also be more useful to analyze FB15K-237 and WN18RR instead of only FB15K.

D2. (W2 ctd.) Authors claim "When a new fact was added into Freebase, it would be added as a pair of reverse triples, denoted explicitly by a special relation reverse_property..." The claim needs justification or citation, it is not clear why it holds with no example of the ontology of FB15k or reference to previous work or sources that created FreeBase.

D3. (W2 ctd.) "There is no prior systematic study with the main objective of assessing the true effectiveness of embedding models in real-world settings" this is not completely true, it has been long known that FB15K has many data leakages and also the same for WN18, that is how WN15K-237 and WN18RR have been introduced in the first place.

D4. (W2 ctd.) Authors claim "Furthermore, the results show it can outperform embedding models, as the FHits@10↑ of AMIE is the highest on FB15k among all methods." If you compare with rotatE (https://arxiv.org/pdf/1902.10197.pdf) this no longer holds based on your reported results for AMIE. Also in table 5, AMIE only has hits@10 results, hits@1 and MRR should be included too + source of purple color is not identified.

9. Candidate for a revision?

Yes

10. Required changes for a revision, if appropriate (labeled R1, R2, R3, etc.).
R1. Include analysis of WN18 too, having the code it should be fairly straight forward.

R2. Include a deeper analysis of FB15K-237 and any possible leakage there. Same for WN18RR.

R3. Update tables that overlap with the other mentioned paper and add the recent embedding models instead of some fairly outdated ones. (Address W1) + Include full results of AMIE.

12. Comments on the revised paper (if appropriate)
The authors did a thorough job addressing my comments with extensive rewrite and added datasets/experiments. The introduction and paper overall becomes a bit long - it would be good for authors to highlight the main messages and reduce some of the detailed analysis/numbers.

13. Recommended decision for the revised paper (if appropriate)
Accept

Reviewer #5

Questions

1. Overall evaluation
Weak Reject

2. Reviewer's confidence
Knowledgeable

3. Novelty
Medium

4. Importance: select all that apply:
SIGMOD attendees will learn something interesting from the paper
The paper contains controversial ideas and/or will generate interesting discussion
The paper is likely to influence other research in the community

5. Summary of contribution (in a few sentences)
Many link prediction methods for knowledge graph completion are evaluated using the FB15K dataset. The paper argues that this dataset is not a good benchmark dataset, and provides supporting evidence.

The paper provides a detailed analysis of the characteristics of the FB15K dataset and identifies characteristics of some of the relations in this dataset that lead to data leakage and thereby training and evaluation. Specifically, the paper identifies that some relations in FB15K (1) have inverse relations in the dataset, or (2) have overlapping relations in the dataset (other relations encoding almost the same information), or (3) are Cartesian product relations that connect all possible (subject, object) pairs. This leads the authors to question the adequacy of FB15K as a benchmark dataset, and to question the algorithmic superiority of newly proposed models compared to early ones like TransE. The paper also suggests the inadequacy of the assumptions made by current evaluation metrics (closed world assumption or local closed world assumption) and highlights the contradiction with one of the main motivating tasks for embedding models which is knowledge graph completion.

The paper provides an interesting analysis of a commonly used dataset, but it is makes overly strong claims while being too narrowly focused on this one dataset.

6. List 3 or more strong points, labeled S1, S2, S3, etc.

S1: Identifies the inverse, overlapped, and Cartesian product relations in FB15K and analyzes their frequency

S2: Provides experiments that suggest a considerable impact of the presence of these types of relations on the training and evaluation process and results

S3: Discusses the variability in the difficulty of predicting different relations in the FB15K dataset and highlights the inadequacy of using the same embedding model to learn all relations

S4: Identifies some of the very easy relations to predict and provides initial experiments for much simpler models performing better than the embedding models on these relations

S5: Timely and topical paper given the interest in graph embedding models

7. List 3 or more weak points, labeled W1, W2, W3, etc.

W1: The paper correctly points out the weaknesses in the FB15K dataset. Generalizing from that to the claim that the paper presents a realistic re-evaluation of the embedding models is not justified.

W2: Focuses exclusively on FB15K, ignoring other datasets commonly used in link prediction work such as WN18 (WordNet), and bigger datasets that are sometimes used such as YAGO10. Once again, generalizing from one dataset is not justified.

W3: Claiming that some type of link prediction tasks are unrealistic based on the way that Freebase was constructed is not justified.

W4: The observation about inverse relations is not new given [27].

W5: The paper unnecessarily focuses on TransE while its analysis and conclusions are not restricted to that model.

8. Detailed evaluation. Number the paragraphs (D1, D2, D3, etc.)

D1: The paper does a good job of presenting the shortcomings of the FB15K dataset. Thus, it shows that the current evaluation protocols and benchmark dataset are inadequate. However, the paper makes general claims about re-evaluation of embedding models, and about fundamental weaknesses in these models. These generalized claims are not justified by experiments on one dataset.

D2: The paper should consider other datasets such as WN18 and YAGO10.

D3: For a fair comparison of the different link prediction methods, the paper should use one software framework that implements all the models, loss functions, and optimizers so that the results are truly comparable.

D4: The paper can include theoretical or experimental analysis to drill down on the deficiencies of different link prediction methods.

D5: Claiming that some type of link prediction tasks are unrealistic based on the way that Freebase was constructed is not justified. For example, the paper claims that predicting an inverse relation is unrealistic because in Freebase, whenever a triple with some relation is added to the knowledge base, the inverse triple is also added. This is indeed the way that Freebase was constructed, but not all knowledge graphs have to be constructed that way. Many knowledge bases have meaningful inverse relations, and these relations are not always complete. Thus, predicting an inverse relation can often be required. A good relational machine learning model should be able to learn that two relations are inverses of each other and accurately predict the inverse if it is missing. Thus, a more nuanced discussion of this point is required.

D6: The paper focuses in the presentation on TransE. This leads the reader to infer that the analysis and conclusions are somehow more applicable to TransE than to other embedding models. There is no reason for this focus, since the analysis and conclusions are equally applicable to all embedding models. Thus, the paper should introduce all models not just TransE and its variants, and should present the conclusions as applicable to all models. The main classes of models are translational (e.g., TransE), multiplicative (e.g., RESCAL), and deep learning (e.g., ConvE). Note that TransE, while popular, was shown to have a theoretical limitation in that it is not universal (https://www.aaai.org/GuideBook2018/16900-72310-GB.pdf)

D7: In the introduction, the paper states that "A relation r, also represented as a vector r, is an operation on h and t, e.g., translation [4], which is a geometric transformation between the head and tail entities in the embedding space." This definition of r applies to TransE but not to other models. For example, in multiplicative models such as RESCAL, r is a feature representing the weights of interactions of the latent features.

D8: The introduction states that TransE [4] is the first translational model. RESCAL [22] was actually published two years prior to TransE.

9. Candidate for a revision?

Yes

10. Required changes for a revision, if appropriate (labeled R1, R2, R3, etc.).

Please address D1-D8. It is important to run experiments on other datasets such as WN18 and/or YAGO10 to see if the observed behavior of different techniques holds beyond FB15K.

12. Comments on the revised paper (if appropriate)

The authors have mostly addressed my revision requests.

13. Recommended decision for the revised paper (if appropriate)

Accept