AI6122 Literature Review on Knowledge Graph Applications

Teng Guang Way, G2102434F

ACM Reference Format:
. 2022. AI6122 Literature Review on Knowledge Graph Applications: Teng Guang Way, G2102434F. In *Proceedings of Data (Text).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

For this assignment, papers on the latest topic of Knowledge Graph Applications have been read and researched. The understandings were expressed through the report assignment in the following ***STARS*** flow format:
1. ***S****ituation*: What is the problem/motivation? Why did the author publish this paper?
2. ***T****asks*: What task or idea did the author mentioned to solve the *Situation*/problem or satisfy the motivation?
3. ***A****ctions*:: What actions did the author performed to complete the *Tasks*?
4. ***R****esults*: What was the reported result and effectiveness of the *Actions* according to the paper?
5. ***S****ummary of thoughts*: After reading the paper and understanding its *STAR*, what are my personal thoughts?

Three papers related to this subject were selected for reading from the SIGPR22 papers. The titles of the papers are mentioned on the Reference page. [1–3]

2 META-KNOWLEDGE TRANSFER FOR INDUCTIVE KNOWLEDGE GRAPH EMBEDDING. [1]

2.0.1 Situation. In this paper, the author introduced the trend of Knowledge graph embedding (KGE) methods being proposed to embed entities and relations of a KG into vector spaces. Such embedding methods provide easier solutions for in-KG tasks(e.g links prediction) and out-of KG tasks(Question Answering). However existing KGE methods are not applicable to inductive settings, where the task involves evaluating entities unseen during training of the KGE model.

2.0.2 Tasks. To overcome such problems, the authors proposed the idea of Inductive Knowledge Graph embeddings being designed and modeled to produce entity embeddings with triples that contain entities that are not seen during training. The authors raise the task involving identifying the common neighboring structures of the unseen entities that were observed during training and then evaluating how they are related to certain kinds of seen entities. Then, based on that observation and evaluation, the entity embedding of the unseen entity can be approximated as something similar to what was seen during training using information that describes the common neighboring structures. The produced Inductive KG embedding is used to tackle the problem of handling in-KG tasks and out-of KG tasks in an inductive setting.

2.0.3 Actions. The authors of the paper proposed a modified version of the KGE model to achieve an inductive knowledge graph embedding model. They named the model MorsE which stands for Meta-Knowledge Transfer for Inductive Knowledge Graph Embedding. The model is designed to produce high-quality embeddings for new entities in the inductive setting, by learning transferable meta-knowledge rather than the embeddings of the entities seen during training like in conventional KGE. The design focus on the development of the two modules namely Entity Initializer and GNN Modulators, which aim to capture type-level information of entities and instance-level information respectively. Using the neighbor structural information of entities from training, the model learns to identify new unseen entities which have neighboring structural information it witnessed in the past. The authors implemented the model keeping the core tasks(In-KG/out-of-KG Tasks in Inductive Settings) in mind. They tested the performance of their model on link prediction and QA answering using public datasets and compare their model competency relative to other state-of-art related works (referred to as baseline models in the paper) to evaluate the core tasks.

2.0.4 Results. The authors reported outperforming state-of-the-art results that their models can tackle the core tasks on stimulated inductive link prediction and QA answering tasks, beating most other existing baseline models for the majority of the public datasets used as benchmarks for comparison. The model was also highlighted to be more robust to the sparsity of the target KG used for testing than the other baselines. The authors ended the paper by highlighting their intention to explore more inductive settings for tasks related to Knowledge Graph for further evaluation on more applications.

2.0.5 Summary of thoughts. The authors developed the KGE model for inductive settings and benchmarked it against common datasets used for training and testing highlighting how it “outperforms existing state-of-the-art models”. However successful testing of the model’s ability to generalize to such a dataset does not directly evidently proves that the model can handle real-life practical application problem better than other models as during real life, the target KG can vary greater or more unexpectedly than a very conventionally used well-processed, and understood public dataset used for such benchmarking evaluation for many years. The quality of the datasets may not be sufficiently good enough to represent real life. The author mentioned the intention to explore more inductive settings tasks for testing. I feel that rather than applying the developed model to similar inductive settings used in the paper for benchmarking, which are artificially stimulated inductive settings,

Text, Management, Individual Assignment

https://doi.org/10.1145/nnnnnnn.nnnnnnn

the authors may consider applying their model to more realistic inductive settings if they are determined to prove the model's definite superiority over other states of the art model. For example, data processes the testing set more tediously by having human to add additional annotations, links, and new entities manually to make the dataset dense by TREC Pooling referring to Paper 3[3].

3 INCORPORATING CONTEXT GRAPH WITH LOGICAL REASONING FOR INDUCTIVE RELATION PREDICTION[2]

3.0.1 Situation. The construction of KGs can be an extremely challenging process. This resort to the reliance on incomplete KGs in applications which further weakens the performance of downstream applications. Many methods to mine missing triples in KGs have been proposed. Most of these models are designed for transductive settings where the entities to be predicted must be seen during training. However, inductive settings are closer to real-life scenarios where new entities are constantly emerging and such models and such models that rely on a transductive setting where the entities are limited are not as applicable.

3.0.2 Tasks. In this paper, the authors propose a novel model, named ConGLR, or Context Graph with Logical Reasoning for Inductive relation prediction, which satisfies modeling and reasoning level requirements of inductive relation predictions. The knowledge graph embedding model also uses relation embedding for the first time in addition to entity embedding. In addition, the model attempted to use neural calculation and logical reasoning cooperatively. (Referred to as the hybrid method in the paper)

3.0.3 Actions. The deployment of the novel model was first performed. The training knowledge graph triples were split into two sets one as the input knowledge graph, and the other as a triples set used for prediction. The procedure to deploy the model can be described with Algorithm 1 in the paper.

At the end of the training, a set of FOL rules are produced each with a confidence score. During testing, the test set is also split into an Inductive graph and a list of triples for prediction. All entities within the test set are not seen in the training set, however, the relation paths and relations were seen. The goal of the model in the testing phrase was designed to predict whether a target triple is valid or not (Positive triple or Negative triple). These were done by using the relation paths between the target head entity and the target tail entity and evaluating all possible relations from the rule set which was learned from the model. Each of the rule confidence can be calculated by equations 14 and 15 in the paper with the relation embeddings and relation paths embeddings learned during training using the context graph. Each rule can correspond a relation path to a relation with a certain confidence score that was learned during training. With a set of relation paths for a given triple, the target relation can be predicted by evaluating the learned FOL rules associated with each of the relation paths in the relation paths set. Some example output rules were also provided in Table 8 in the paper for reference. The inference of the FOL rules based on the relation paths was used to evaluate the final confidence score to assess if the triple in testing is a valid triple in an inductive setting.

3.0.4 Results. The trained CONGLR model is then compared against six state-of-the-art models using 12 datasets of different sizes derived from subsets of the WIN18RR, FB15K-237, and NELL-995 datasets with the two micro metrics Area under precision-recall curve and Hits@10. The model demonstrates state-of-the-art performance as it consistently performs best for most of the datasets used for benchmarking. The author stressed that the benchmark result "demonstrates the superiority of their method over state-of-the-art baselines".

3.0.5 Summary of thoughts. The modeling of ConGLR was evaluated with a wide range of datasets. However, like in the case of Paper 1[1]. Most of these datasets were well studied, and the model already has some knowledge of the nature prior to the development of these models. In real life inductive settings, new entities and triples emerge constantly. Therefore scoring 1 to 2% better than the other state of art models for Hits@10 does not accurately prove that the model is in any way better than the other models during real-life applications. The ranking may be unstable should there be a change in the dataset.[3] Therefore the model ranking might not be reflected accurately based on the authors' benchmarking. To provide superiority over other state-of-the-art models, a superior dataset may be designed by TREC pooling such as in the case of Paper 3[3] for benchmarking all the other models and a more thorough investigation is required. However, from the benchmarking, it was clear that the model performance was sufficient to be considered a valid novel KG completion model.

4 RE-THINKING KNOWLEDGE GRAPH COMPLETE EVALUATION FROM AN INFORMATION RETRIEVAL PERSPECTIVE. [3]

Notes: In this paper, the summarizing strategy involves splitting the report content into two *STAR* contents and one *Summary of Thoughts*. The first *STAR* centralized around a situation regarding the impact the test dataset's sparsity may have on the evaluation of KGC through entity predictions. The second *STAR* centralized around a situation motivated from the previous *STAR* result, and the possibility of whether an alternative metric, called macro metrics used in IR evaluation could be used as a better evaluation metric. The review ends with a summary of my personal thoughts after reading the paper. The overall flow of reporting is described in Figure 1.

4.0.1 First STAR- Situation. In this paper, the conventional Knowledge Graph Completion evaluation protocol was highlighted to be flawed as the evaluations are based on extremely sparse labels with micro metrics. For each triple question, there could be multiple factual answers. For example, given a triple question (Trump, visited, ?), there are multiple factual answers like China, Japan, India, Singapore, United Kingdom, etc, just like in the case of an IR evaluation of a query. Sparse datasets like FB15K-237 may only contain the knowledge to provide one of the above answers for that specific triple. Therefore, a sparse dataset like FB15k-237 may be flawed for direct evaluation testing of the KGC model performance, as a KGC model that uses a higher quality testing dataset which

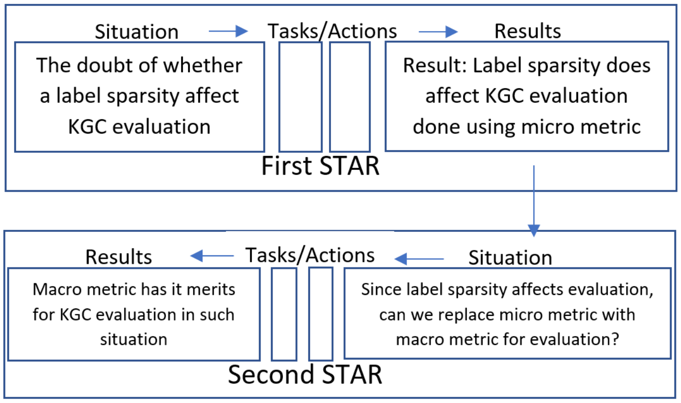
Situation
Tasks/Actions
Results
The doubt of whether a label sparsity affect KGC evaluation
Result: Label sparsity does affect KGC evaluation done using micro metric
First STAR
Results
Tasks/Actions
Situation
Macro metric has it merits for KGC evaluation in such situation
Since label sparsity affects evaluation, can we replace micro metric with macro metric for evaluation?
Second STAR

Figure 1: Report process flow for this Paper

is denser may result in a different inference result of the model's performance.

4.0.2 First STAR- Tasks. In this paper, the authors specifically address if label sparsity affects KGC evaluation. The author then proposed the idea to use 2 datasets, one that is sparse and the other dense, for comparison of the performance metrics (micro metrics) changes with multiple KGC models.

4.0.3 First STAR- Actions. To verify the above research question, the authors perform the following procedures to obtain the above 2 datasets required for comparison.

For the flawed sparse dataset, the authors used the dataset commonly used for evaluating KGC entity ranking by simply sampling a small subset of the FB15k-237 dataset. (Referred to as FB-Test-S in the paper)

For the dense dataset, based on inspiration from the TREC pooling method commonly used in IR evaluation to handle label incompleteness, the authors process tediously on the small subset named FB-Test-S, which initially contains 1023 test triples. Each of the triple questions of the small subset is then used to generate more entity answers by filtering and collecting the predicted triples from six diverse well-tuned baseline KGC modeling systems. These collected answer entities are then pooled into the small subset as new entity answers of their respective triple question to form a denser dataset. An additional manual labor annotating process then further adds more answer entities. The final produced dense dataset is coined FB-Test-S-C.

With the 2 datasets, the authors conduct experiments to assess whether if label sparsity affects KGC evaluations by using the micro metrics to report the per-answer performance of KGC models.

4.0.4 First STAR- Results. The authors noticed that by pooling more answer entities to form a denser set, the correlation of system rankings between the sparse and dense dataset drops per Figure 4 in the paper. Therefore, the effects of data sparsity are not negligible.

4.0.5 Second STAR- Situation. To evaluate a KGC model effectively, one would want the evaluation method to be based solely on discriminating between better models and lousier models. Therefore, the data sparsity effect on KGC evaluations proven in First Star-Results is undesirable as it results in incorrect assessments of the KGC model competency relative to others.

4.0.6 Second STAR- Tasks. The authors highlighted if there is a possibility to negate or make less significant the influence of data sparsity on the evaluation by the introduction of macro metrics to reflect the ranking nature of the KGC task. The paper then set out on tasks to verify if the macro metrics used to evaluate IR systems may differ from the micro metrics used in KGC evaluation and could macro metrics be used for KGC evaluation to reduce the effects of data incompleteness. The discriminative power of the macro metrics was also proposed to be evaluated.

4.0.7 Second STAR- Actions. The correlations of the macro metrics based system ranking between a sparse dataset and dense dataset with varying pooling depth were compared with micro metrics to see if the changes in correlation are less significant. In addition, the authors also tested the data sensitivities by simulating a series of test subsets in different sizes of their original and then evaluating the models using different macro metrics and micro metrics. The authors finally went on to evaluate the discriminative power of system pairs for each metric. The discriminative power is defined as p-values, where a smaller p-value between the two systems means that the better performance system was discriminated against based on its performance as intended and not by other unknown effects.

4.0.8 Second STAR- Results. The introduction of the macro metrics as shown in Figure 5 of the paper does show some decrease in changes in the evaluation correlations between the sparse and the increasingly dense datasets. Macro metrics are also noted to be more stable than micro metrics when the datasets were tested for data sensitivities for the sparse dataset as shown in Figure 7a in the paper. The authors also plotted their result of discriminative power analysis noting that both macro and micro metrics are discriminative to distinguish model performance, however, macro metrics are better to handle incomplete test graphs as it was able to obtain a smaller overall p-value as shown in Figure 8 in the paper. The author concluded that Macro metrics are more stable, discriminative, and less sensitive to label sparsity than micro metrics and can be used together with micro metrics to evaluate the various aspect of KGC models. The authors also suggest the use of TREC-style pooling to deal with label incompleteness problems when one evaluates a model directly with the FB15K-237 dataset triple questions.

4.0.9 Second STAR- Summary of Thoughts. The scope of this paper covered what I felt was lacking in Paper 1[1] and Paper 2[2], more specifically, flaws in the evaluation when one designed a model specifically for evaluation on the well-processed, well-understood, FB15K-237. The paper went very in-depth to explain the evaluation problems in Knowledge Graph Completion modeling. However, I feel that the authors struggle to collate their points neatly due to the overwhelming amount of context to cover. English language may not be their first writing language and some phrases were difficult to comprehend quickly. To read and understand the paper thoroughly, more time and patience are required as compared to the other two papers read in this reading assignment.

REFERENCES

[1] Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. 2022. P1: Meta-Knowledge Transfer for Inductive Knowledge Graph Embedding. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531757

[2] Qika Lin, Jun Liu, Fangzhi Xu, Yudai Pan, Yifan Zhu, Lingling Zhang, and Tianzhe Zhao. 2022. P2: Incorporating Context Graph with Logical Reasoning for Inductive Relation Prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531996

[3] Ying Zhou, Xuanang Chen, Ben He, Zheng Ye, and Le Sun. 2022. P3: Re-thinking Knowledge Graph Completion Evaluation from an Information Retrieval Perspective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3532052