A Simplified Benchmark for Ambiguous Explanations of Knowledge Graph Link Prediction Using Relational Graph Convolutional Networks (Student Abstract)

Nicholas Halliwell,[1] **Fabien Gandon,**[1] **Freddy Lecue**[1, 2]

[1] Inria, Université Côte d'Azur, CNRS, I3S, France
[2] CortAIx, Thales, Montreal, Canada
nicholas.halliwell@inria.fr, fabien.gandon@inria.fr, freddy.lecue@inria.fr

Abstract

Relational Graph Convolutional Networks (RGCNs) are commonly used on Knowledge Graphs (KGs) to perform black box link prediction. Several algorithms have been proposed to explain their predictions. Evaluating performance of explanation methods for link prediction is difficult without ground truth explanations. Furthermore, there can be multiple explanations for a given prediction in a KG. No dataset exists where observations have multiple ground truth explanations to compare against. Additionally, no standard scoring metrics exist to compare predicted explanations against multiple ground truth explanations. We propose and evaluate a method, including a dataset, to benchmark explanation methods on the task of explainable link prediction using RGCNs.

Explainable GCN Link Prediction

Knowledge Graphs represent facts as triples where a *subject* and *object* representing real world entities are linked by some *predicate*. Link prediction methods discover new facts from existing ones. One method is to use graph embeddings, where a function is learned to map each subject, object, and predicate to a low dimensional space. A Relational Graph Convolutional Network (RGCN) (Schlichtkrull et al. 2018) leverages Graph Convolutional Networks (Kipf and Welling 2017) with a scoring function such as DistMult (Yang et al. 2015) as an output layer, returning a probability of the input triple being a fact.

Recent research has been devoted to develop methods to explain the predictions of Graph Neural Networks on link prediction. GNNExplainer (Ying et al. 2019) explains the predictions of any Graph Neural Network, learning a mask over the adjacency matrix to identify the most informative subgraph. ExplaiNE (Kang, Lijffijt, and Bie 2019) quantifies how the predicted probability of a link changes when weakening or removing a link with a neighboring node.

The weakness of these papers is their evaluation of explanation quality due to the lack of available datasets with ground truth explanations. In this work, we propose a method, including a dataset (FrenchRoyalty-200k), to quantitatively evaluate non-unique explanations. We adapt several scoring metrics for use on this dataset, and perform a benchmark comparing state-of-the-art explanation methods.

User-Scored Non-Unique Explanations

In a Knowledge Graph, the available formal semantics allow us to view ground truth explanations as equivalent to computing justification for an entailment. We select an open-source semantic reasoner with rule-tracing capabilities (Corby et al. 2012) to generate ground truth explanations for each defined rule. The explanation associated with a triple consists of all possible triples that triggered each rule. This tracing pinpoints the input triples that caused the generation of a triple we will then try to predict and explain. All resources used and produced in this work are available online.[1]

From DBpedia, we built a Knowledge Graph of the French Royalty focusing on 6 family relationships: *hasSpouse*, with 3 possible explanations, *hasBrother*, with 7 possible explanations, *hasSister*, with 7 possible explanations, *hasGrandparent*, with 6 possible explanations, and *hasChild*, and *hasParent* with 9 possible explanations. We distinguish between two types of rules, logical derivation and partial explanation. We define a *logical derivation rule* as one that is always true, and a *partial explanation rule* as one that is not always true without additional information, such as gender. The logical derivation rules trigger every time their antecedent is matched, and its corresponding triple and logically true explanation are generated. The partial explanation rules trigger only if the triple is already known (asserted or inferred by other rules) and are just adding alternative partial explanations, therefore preventing any false triples from being included in the graph.

We conducted a survey to score each possible explanation rule, allowing us to distinguish explanations that are intuitive from those that are not without relying on any prior assumptions. In total, 42 users responded from 11 different nationalities, consisting of both computer science and non-computer science backgrounds. We normalized the average scores between 0 and 1 for each possible explanation, and round them to the nearest tenth. These user scores are used in the rules and in the benchmark, to penalize unintuitive predicted explanations, and reward intuitive predicted explanations.

[1] https://github.com/halliwelln/multiple-explanations/

Models	Metrics	Predicate						
		Spouse	Brother	Sister	Grandparent	Child	Parent	Full data
RGCN	Accuracy	0.786	0.878	0.826	0.822	0.804	0.8	0.81
GNN Explainer	Generalized Precision	0.071	0.174	0.117	0.129	0.109	0.091	0.11
	Generalized Recall	0.106	0.192	0.142	0.129	0.125	0.102	0.121
	Generalized F_1	0.083	0.18	0.126	0.129	0.114	0.095	0.114
	Max-Jaccard	0.066	0.2	0.151	0.102	0.125	0.12	0.11
ExplaiNE	Generalized Precision	**0.138**	**0.25**	**0.194**	**0.177**	**0.166**	**0.182**	**0.173**
	Generalized Recall	**0.221**	**0.263**	**0.214**	**0.177**	**0.207**	**0.222**	**0.2**
	Generalized F_1	**0.165**	**0.253**	**0.2**	**0.177**	**0.18**	**0.195**	**0.182**
	Max-Jaccard	**0.133**	**0.27**	**0.237**	**0.145**	**0.187**	**0.225**	**0.174**

Table 1: Benchmark results on FrenchRoyalty-200k: Link prediction results for RGCN, and explanation evaluation for GNNExplainer and ExplaiNE on each subset of the full dataset. Highest scores per predicate denoted in bold.

Evaluating RGCN Explanation Quality

The binary precision and recall could be used to measure performance for this task, however, these metrics fail to account for the fact that some explanations can be more intuitive than others to users. Both metrics would give a score of 1 when a predicted explanation exactly matches a ground truth explanation. However, an explanation method could predict an unintuitive explanation, and receive the highest possible evaluation score, potentially misleading practitioners into thinking the predicted explanation is of high quality. Therefore, scoring metrics used for this task must compare a predicted explanation to all possible explanations, and account for the fact that explanations have different degrees of relevance.

We adapted the generalized precision and generalized recall (Kekäläinen and Järvelin 2002) to the context of link prediction on Knowledge Graphs, using the user scores as relevance weights for each explanation. Furthermore, we proposed the use of the max-Jaccard across all possible explanations for a given triple. The max-Jaccard score measures if the explanation method is able to accurately predict one of the possible explanations to choose from. The generalized precision and generalized recall measure if the predicted explanation was given a high intuitive score assigned by users. Both metrics prevent an explanation method from only predicting low scored, unintuitive explanations, and receiving a high score. Lastly, the generalized F_1 provides an overview of performance on the generalized precision and recall.

Table 1 breaks down the results on the FrenchRoyalty-200k dataset. We filter the results on the full data for each predicate and compare performance metrics to each predicate subset. For example, the Spouse column of Table 1 reports the benchmark performance of all input triples with the *hasSpouse* predicate from an RGCN trained on the full data, hence the RGCN is exposed to all possible predicates.

We find ExplaiNE outperforms GNNExplainer on all subsets, across all metrics. Upon examining the errors, we find that both explanation methods do not always attempt to predict explanations with the highest user scores. We find that both explanation methods instead predict explanations with the most common user scores.

Conclusion

We propose a method, including a dataset (FrenchRoyalty-200k), to benchmark explanation methods when there are multiple ground truths to consider. We adapt several scoring metrics to account for differences in explanation simplicity. We benchmark two state-of-the-art explanation methods, ExplaiNE and GNNExplainer, using the proposed dataset and scoring metrics.

References

Corby, O.; Gaignard, A.; Faron Zucker, C.; and Montagnat, J. 2012. KGRAM Versatile Inference and Query Engine for the Web of Linked Data. In *IEEE/WIC/ACM Int. Conference on Web Intelligence*, 1–8. Macao, China.

Kang, B.; Lijffijt, J.; and Bie, T. D. 2019. ExplaiNE: An Approach for Explaining Network Embedding-based Link Predictions. *CoRR*, abs/1904.12694.

Kekäläinen, J.; and Järvelin, K. 2002. Using graded relevance assessments in IR evaluation. *J. Assoc. Inf. Sci. Technol.*, 53(13): 1120–1129.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Int. Conf. on Learning Representations, ICLR*.

Schlichtkrull, M. S.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling Relational Data with Graph Convolutional Networks. In *European Semantic Web Conference, ESWC*.

Yang, B.; Yih, W.; He, X.; Gao, J.; and Deng, L. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *3rd International Conference on Learning Representations, ICLR*.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems*.