Interpreting Link Prediction on Knowledge Graphs

Andrea Rossi
Supervised by proff. Paolo Merialdo and Donatella Firmani

Roma Tre University, Rome, Italy andrea.rossi3@uniroma3.it

Abstract. Link Prediction (LP) on Knowledge Graphs (KGs) has recently become a sparkling research topic, benefiting from the explosion of machine learning techniques. Several relation-learning models are published every year, mostly relying on KG embeddings. So far, however, not much has been done to interpret the features they learn and predict, and the circumstances that allow them to achieve satisfactory performances. Our research aims at opening the black box of LP models, trying to explain their behaviors. In this work we first discuss the current limitations of LP benchmarks, showing how the use of global metrics on largely skewed datasets hinders our understanding of these models; we then report the main takeaways from our recent comparative analysis of state-of-the-art LP models [3], identifying the most influential structural features of the graph for predictive effectiveness.

Keywords: Knowledge Graphs · Knowledge Graph Embeddings · Link Prediction.

1 Introduction

Knowledge Graphs (KGs) model data as nodes linked by labeled edges. In a KG nodes represent *entities*; each edge connects a *head* entity to a *tail* entity through the *relation* specified by its label, resulting in a *fact*. KGs are employed in several domains, ranging from question answering to information retrieval and content-based recommendation. All KGs tend to suffer from incompleteness; Link Prediction (LP) tackles this issue by leveraging the known facts to infer the missing ones. LP research has been largely influenced by the recent advancements in machine learning; most LP models nowadays map the KG elements into vectors dubbed KG embeddings, learned automatically based on scoring functions that estimate the plausibility of the training facts. For instance, $\langle$*Barack Obama*, *born_in*, *Honolulu*$\rangle$ is expected to yield a better score than $\langle$*Barack Obama*, *born_in*, *Beijing*$\rangle$. In this framework, predicting the tail of an incomplete fact $\langle h, r, ?\rangle$ amounts to finding the entity that results in the best

score when used as its tail. Head prediction is performed analogously. Despite the popularity of LP techniques, their strengths, weaknesses and limitations are still unknown, and the graph structural features driving predictions have been hardly investigated. Roughly speaking, we still do not really know what makes a fact easy or hard to learn and predict, and whether the corresponding predictions can be trusted or not. Our research focuses on interpreting the behavior of LP models, aiming at providing tools for explaining their predictions. As a first step towards this goal, in this paper we summarize our findings on the limitations of current benchmarks in showing how different papers in LP literature fit together. Then, we report the main results of our comparative analysis. Finally, we discuss our research plans for building interpretable methods for LP.

2 Related Works

Works related to ours are mostly *meta-analyses* focusing on specific LP methodologies. These works tend to address very specific hypotheses to interpret LP behaviours, and run experiments on a few selected models to verify them. For instance, the authors of [5] study geometrical properties of the embedding vectors, measuring their Alignment To Mean (ATM) and conicity. They show that models that operate by adding embeddings tend to learn significantly sparser vectors than the ones multiplying them; in the latter, higher conicity also seems to correlate to better performances. The work of [7] points out that the current evaluation practices just ensure that models prioritize correct answers over wrong ones on test facts. In this way, only questions that do have an answer are taken into account. The authors argue that this approach is more akin to question answering than LP, and propose a novel measure that includes questions with no correct answers (e.g. nonsensical questions, such as $\langle Apple, gender, ?\rangle$).

Finally, we acknowledge that LP is also being researched on standard graphs. In this scenario edges are usually *non-labeled*, so modeling relations is unnecessary: this makes it a related but ultimately very different task from LP on KGs.

3 Interpreting Link Prediction Results

Since the seminal work of [1], dozens of models have been developed in just a few years (see [6] for a survey). We argue that a crucial step towards interpreting these models lies in providing informative evaluation practices. We briefly highlight the most prominent limitations of current LP evaluation practices and benchmarks, and then summarize our main findings when comparing LP models.

3.1 Benchmark Limitations

All the currently most popular LP datasets have been generated by sampling facts from a KG and splitting them uniformly at random into a training, a validation and a test set. As a side effect, in such datasets the number of mentions

of both entities and relations display significantly skewed distributions: less than 15% entities can be featured in more than 80% training facts. Furthermore, since the training and test set are random splits from the same original sample, the most mentioned entities in training are largely over-represented in testing too.

Since current evaluation practices rely almost solely on global metrics (e.g. *Hits@K*, *Mean Rank*, *Mean Reciprocal Rank*) over the entire test set, LP models can exhibit good performances, in proportion, by just learning to predict the most mentioned entities, while ignoring the others. In FB15k [1], one of the first LP datasets and a *de facto* standard, "United States" is both the most mentioned entity and the most common answer to relation "nationality"; in this setting, a model can obtain decent results by just learning to predict U.S. citizens only.

In our work [4] we have observed experimentally that LP models are indeed subject to this issue to some extent, as they achieve better performances when dealing with entities with more training mentions.

3.2 Comparative Analysis of Models

As mentioned above, relying exclusively on global metrics hides any variations in predictive performances across different portions of the dataset. This makes it difficult to analyze the conditions that facilitate or hinder predictions. We also acknowledge the difficulty of coming up with new datasets with different structural properties and good semantic consistency. In order to mitigate this issue, we have proposed a set of evaluation practices going beyond what is available in literature, taking into account structural properties of individual facts and entities. We have run an extensive comparative analysis [3] on a set of 16 models representative for the most successful architectures applied to LP; we have included an additional rule-based LP model as a baseline. We have trained and fine-tuned all models on the 5 most popular LP datasets, extracting fine-grained results with the predictions yielded by each model for each test fact. The full list of featured datasets and models can be found at our repository.

We have used these results to investigate the graph structural features that make facts easier to learn and predict, searching for the strongest correlations with the predictive performances of the models. The structural features we found most influential to predictive performances are the number of *peers* and the support provided by *paths*, as discussed below.

Given a prediction, its target is the entity to predict, and its *target peers* are its correct alternatives, i.e. forming a fact belonging to the dataset. When a prediction has too many target peers, LP models seem to get confused: as they try to optimize the embeddings for too many correct answers, they also let many incorrect ones in, leading to a rapid decrease in performances. In a specular way, the known entity in the prediction is called its source, and its correct alternatives are dubbed *source peers*. Source peers seem to facilitate predictions, leading to better performances; this can be explained by interpreting them as specific examples that enable analogical reasoning.

https://github.com/merialdo/research.lpca

In a graph, *paths* are chained sequences of facts. Given a fact, the paths connecting its head and tail can provide useful patterns for prediction, e.g. ⟨*Barack Obama*, *born_in*, *Honolulu* ⟩ and ⟨ *Honolulu*, *located_in*, *USA*⟩ can be useful to predict ⟨*Barack Obama*, *nationality*, *USA*⟩. In our work [3] propose a novel *RPS* measure estimating the *Relational Path Support* of any test fact. *RPS* uses TF-IDF vectors to assess the similarity between the paths co-occurring with the specific test fact and the ones usually co-occurring with the other facts that feature the same relation. We have observed that most LP models, despite just training on individual facts, are able to leverage longer-range dependencies to some extent, as higher *RPS* values always correspond to far better performances.

4 Research plan and concluding remarks

Being aware of the behaviours of current LP methods is vital to identify their weaknesses, and to ultimately build more robust, trustworthy systems.

In order to open the black box of LP systems, we aim at building a full-fledged explainability framework: given a prediction of a model, our framework would yield as an explanation the training facts that have been most influential for it. In the categorization of [2], this amounts to a *post-hoc local explanation*. The framework we aim to create should be agnostic to the architecture of the LP model to explain, thus being applicable to a set of systems as wide as possible.

As a matter of fact, explaining the predictions provided by LP models is still an open problem. To the best of our knowledge, the only technique proposed so far is [8]; this approach, however, still displays severe limitations, as it just searches for meaningful paths connecting the head and tail, without any evidence that such paths have actually been instrumental to perform the prediction.

Our research plan also includes the development of more balanced and insightful benchmarking workloads, in terms of both datasets and metrics.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS (2013)
2. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. CACM (2019)
3. Rossi, A., Firmani, D., Matinata, A., Merialdo, P., Barbosa, D.: Knowledge graph embedding for link prediction: A comparative analysis (2020)
4. Rossi, A., Matinata, A.: Knowledge graph embeddings: Are relation-learning models learning relations? In: PIE (2020)
5. Sharma, A., Talukdar, P., et al.: Towards understanding the geometry of knowledge graph embeddings. In: ACL (2018)
6. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. TKDE (2017)
7. Wang, Y., Ruffinelli, D., Gemulla, R., Broscheit, S., Meilicke, C.: On evaluating embedding models for knowledge base completion. In: RepL4NLP@ACL (2019)
8. Zhang, W., Paudel, B., Zhang, W., Bernstein, A., Chen, H.: Interaction embeddings for prediction and explanation in knowledge graphs. In: WSDM (2019)