Comparison of Symbolic & Sub-Symbolic Approaches for Knowledge Graph Completion

Lars Joormann

February 20th, 2022

1 Introduction

My master thesis will be about comparing symbolic and sub-symbolic approaches for Knowledge Graph Completion. I want to compare prediction results between these two approaches to determine whether there are subsets of the datasets, commonly used in this research area, where one approach significantly outperforms the other. If that is the case I plan to try to identify patterns behind this behaviour. In the following I specify my goals and workplan for the thesis.

2 Background

Knowledge Graph Completion considers Knowledge Bases as graphs. In this graph nodes represent entities while the link between two nodes represents their relation. Therefore the graph can be expressed as a set of facts, also called triples: (h, r, t) with h being the head entity, t being the tail entity and r being the relation between those two. The goal of Knowledge Graph Completion, or sometimes also referred to as Link Prediction, is to find missing relations in the Knowledge Graph. [2]

2.1 Symbolic Approaches

To solve the problem at hand, symbolic approaches "[...] learn and apply an explicit symbolic representation of [the] patterns[.]" found in the Knowledge Graphs [3]. One example for such a model is AnyBURL. This model tries to learn rules from the Knowledge Graph. It does so by generating rules from randomly sampled paths of the length n. If the generated rule satisfies the quality criteria it stores the rule. The n starts with a small value to first cover the shorter paths. The model searches rules for this n within a set time frame. After the time frame n is increased by 1 if a saturation of rules of the length $n - 1$ is reached. If not the next time frame searches for rules with the same length n without increasing it. In the end the model returns a set of rules which can be used to predict prior unseen data by checking which rules fire for the new triple to calculate a confidence score based on them. [4]

2.2 Sub-Symbolic Approaches

Another approach to solve the same problem is the sub-symbolic approach. Here latent representations, also referred to as embeddings, are learned for every entity and relation. These representations then get processed, for example through vector operations, to result in a final score which expresses how likely a triple is to be true. [2] An example for such a model is TransE. This model interprets the relation as the difference between the embedding vector of the head and tail entity. [1]

3 Goals and Work Plan

Deepen my Understanding of Knowledge Graph Completion. As a first goal of my thesis I would deepen my understanding of Knowledge Graph Completion. Here I would mostly focus on studying symbolic and sub-symbolic approaches in the context of Knowledge Graph Completion.

Identify Triples where one Approach outperforms the Other. The next goal then would be to start the comparison between the two kinds of approaches by picking two or more models for each approach and comparing the cases where one approach performs better than the other.

Try to Detect a Patterns between these Groups of Triples. With the comparison from above I would then try to find patterns between these two groups. Which I can then in return use to create subsets of the test dataset(s). These subsets could be used to test models for specific vulnerabilities.

(Optional, if possible) Create a Synthetic Dataset. If all of the above works as planned it might also be possible to create a synthetic dataset. This dataset would have the property that symbolic approaches can learn/ predict it at ease while sub-symbolic approaches would have a harder time making good predictions.

Workplan. To achieve the above listed goals I would first run experiments using the same dataset for one symbolic and one sub-symbolic model. Since sub-symbolic models get initialized with random weights, their predictions change between experiments. To nullify this random factor I would run the experiment for the sub-symbolic model multiple times and use the average of these results for my comparison.

With these results I would then compare which triples are best predicted by which model. To verify that the model then properly represents the approach and that it is not only model specific I would repeat the experiments using two other models, again one from each approach. This will result in three groups of triples. The triples better predicted by symbolic approaches, the ones better predicted by sub-symbolic approaches and the triples which were equivalently predicted by both approaches. If needed this could also be split up more granular.

The next step then would be to analyse these results in order to determine why a specific triple or a group of triples is easier to predict for one of the approaches. Here I would have a look into whether the following factors have an influence:

- relation class (1-1, 1-N, N-1, N-M)
- entity/ relation frequency in the training data
- frequency of similar entities in the training data
- amount of rules (from AnyBURL) firing for the specific test triple

To analyse these factors I am currently thinking about creating a dataset which will include all test triples, the approach which predicts these triple best and the factors above for which I will where possible automatically label the data and where it is not I plan on doing it manually. The list of factors might expand during my research. The dataset will then be used to explore the data by using descriptive statistics e.g. calculating correlations and creating appropriate graphs.

Based on the outcome of this I plan to create subsets of test data which test for a specific vulnerability. For example if one of the outcomes is that multiple sub-symbolic models have problems predicting triples with a specific relation class, a good test subset could be one only containing triples from this class.

If everything goes as planned I will afterwards look into expanding these subsets into a synthetic dataset.

References

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[2] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning Structured Embeddings of Knowledge Bases. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 301–306, August 2011.

[3] Christian Meilicke, Melisachew Wudage Chekol, Manuel Fink, and Heiner Stuckenschmidt. Reinforced Anytime Bottom Up Rule Learning for Knowledge Graph Completion. April 2020.

[4] Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. Anytime Bottom-Up Rule Learning for Knowledge Graph Completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3137–3143. International Joint Conferences on Artificial Intelligence Organization, August 2019.