Embedding based Link Prediction for Knowledge Graph Completion

Russa Biswas[1]

Abstract. This thesis proposes a novel Knowledge Graph (KG) embedding model for Link Prediction (LP) for Knowledge Graph Completion (KGC). The missing links in a KG are predicted based on the existing contextual information as well as textual entity descriptions. The model outperforms the state-of-the-art (SOTA) model DKRL for FB15k and FB15k-237 datasets.

1 INTRODUCTION

KGs are large networks of real world entities and relationships between them. The facts are represented as a triple $< h, r, t >$, where h and t are the head and tail entities respectively and r represents the relation between them. Despite the huge amounts of relational data, KGs are sparse and often incomplete as the links between the entities are missing. Furthermore, different KGs have information about the same real world entities but the fact that these entities in different KGs are same is missing.

LP is a fundamental task of KGC that aims to estimate the likelihood of the existence of links between entities based on the current observed structure of the KG. LP task can be performed across different KGs to predict the missing links between two same entities across KGs and is also known as Entity Alignment. This thesis focuses on the KGC task based on predicting the missing links within the KG as well as across multiple KGs.

Due to the high complexity of the graph mining algorithms, the latent representation of a KG is learned into a low dimensional space. To-date many algorithms are proposed to learn the embeddings of the entities and relations into the same vector space. However, none of the SOTA models consider the contextual information of the KGs along with the textual entity descriptions to learn the latent representation for the task of LP within the KG. This thesis focuses on proposing a model which takes the above described features into account and performs the task of LP i.e., head, tail prediction as well as triple classification. On the other hand, due to the structural differences amongst multiple KGs, their embedding spaces also exhibit different characteristics. Therefore, for the entity alignment task, these different vector spaces generated for different KGs are to be aligned to a single space to predict the missing links between the same entities across different KGs.

2 STATE OF THE ART

Link Prediction. So far, different KG embedding techniques have been proposed which can be categorized as translation based models, semantic matching models, models incorporating entity types, models incorporating relation paths, models using logical rules, models with temporal information, models using graph structures, and models incorporating information represented in literals and a detailed description of them are provided in [4], and [7]. Amonsgst them, he translational model [2] use scoring function based on distance and the translation is carried out with the help of a relation. GAKE [6] considers the contextual information by generating paths starting from an entity. On the other hand, DKRL [11] incorporates textual entity descriptions in the embedding model and uses TransE as the base model.

The textual entity descriptions present in the KGs provide information about the entity which might not be available otherwise in the KG. Also, the paths originating from an entity provide the structural contextual information about the neighboring entities. Therefore, in this thesis, paths and entity descriptions are modeled together to learn the embeddings of entities and relations for LP.

Entity Alignment. Entity Alignment is the task of aligning the same entities across different KGs. To do so, several embedding based methods have been proposed, in which a unified embedding space is learned using a set of already aligned entities and triples. A detailed description of these models for entity alignment is provided in [1]. The challenges of these models are: **(i)** They are supervised and require a set of aligned entities or triples as seeds for training. **(ii)** Some of the models require all the relations to be aligned between the KGs. However, in case of heterogeneous KGs which consist of different sets of relations, it is a challenging task to have a pre-aligned set of relations. **(iii)** The methods lack proper mechanisms to handle multi-valued relations. This thesis proposes an entity alignment model for heterogeneous KGs with multi-valued relations based on the unsupervised approach, i.e. without pre-aligned seeds for training.

3 RESEARCH QUESTIONS AND CONTRIBUTIONS

This section discusses the research questions and the corresponding contributions to address the challenges.

- *RQ1: Given an entity and a relation pair, how to predict the missing entity in a triple?*
  - The head or tail entity in a triple $< h, r, ? >$ or $<?, r, t >$ is predicted by defining a mapping function $\psi : E \times R \times E \rightarrow R$, where E and R are the set of entities and relations in the KG. A score is assigned to each triple where the higher the score of the triple indicates the more likely to be true.
- *RQ2: How to identify whether a given triple is valid or not?*
  - This is a triple classification task, in which a binary classifier is trained to identify whether a given triple is false (0) or true (1).

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe Institute of Technology, Institute AIFB, Germany email: russa.biswas@fiz-karlsruhe.de

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 69–70, Santiago de Compostela, Spain, August 29-30, 2020.

- *RQ3: How to predict the type information for an entity in a KG?*
  - Entity typing or Entity Classification is the process of assigning a type to an entity. To do so, different structural and literal information have been exploited to train a multi-label classification model for fine-grained entity typing.
- *RQ4: How to align the different embedding spaces of the KGs into a unified vector space to identify the* `owl:sameAs` *links?*
  - To align two different KG embedding spaces X and Y, a translation function τ coupled with a rotation function θ is introduced. The `owl:sameAs` links are then to be determined by vector similarity.

Therefore, the main contributions of this thesis are:

- A novel KG embedding model exploiting the structural as well as the textual entity descriptions in the KGs for head and tail prediction as well as triple classification.
- A neural network based multi-label hierarchical classification model for fine-grained entity typing using different features in the KG such as text and images along with the structural information.
- A novel translational model to align the different KG embedding spaces to identify the `owl:sameAs` links across multiple KGs.

4 LINK PREDICTION

To encapsulate the contextual information, random walks of 4 hops are generated starting from each entity in the KG. Predicate Frequency Inverse Triple Frequency (PF-ITF) [8] is used to identify the important relations for each entity. A sequence-to-sequence (seq2seq) learning based encoder-decoder model [10] is adapted to learn the representation of the path vectors in the KGs as shown in Figure 1. Given a path sequence, which is a combination of entities and the relations between them, such as $\{e_1, r_1, e_2, r_2, ..., e_n\}$, the input to the encoder is the corresponding embeddings (computed using TransE). These embeddings are passed through an attention based Bi-directional GRU which encapsulates the information for all input elements and compresses them into a context vector which is then passed through the decoder. A scaled dot product is employed as the attention mechanism. The representation of the textual entity descriptions is obtained using SBERT [9], followed by the same encoder-decoder model. ConvE [5] is used as a base model for scoring. For triple classification, the vectors are passed through a CNN model. Both triple classification and head/tail prediction are evaluated for FB15k and FB15k-237 datasets and the model outperforms the SOTA model DKRL as depicted in Table1. For the entity typing task, a multi-label CNN model is to be built.

5 FUTURE WORK

Entity Alignment. This task is yet to be addressed in this thesis. However, the basic idea is to adapt MUSE [3] which is an unsupervised multi-lingual word embedding alignment model to the KG alignment. A translation function coupled with a rotational function is to be used to align the related entities from different KGs. The same or related entities in different KGs will have overlapping information which could be exploited in an unsupervised manner.

ACKNOWLEDGEMENTS

This thesis is supervised by Prof. Dr. Harald Sack and Dr. Mehwish Alam.

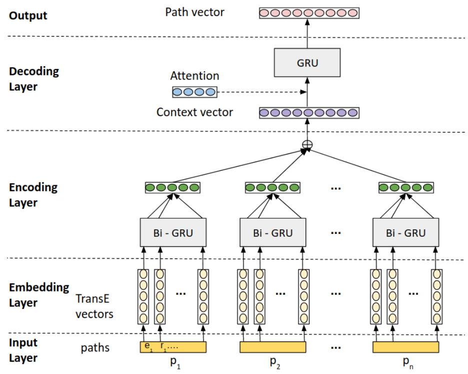

Figure 1. Encoder-Decoder Architecture

Table 1. Results on LP with FB15k and FB15k-237 datasets.

Models	MR	MRR	Hits@1	Hits@3	Hits@10
FB15k					
DKRL	85.5	0.311	0.192	0.359	0.548
Our model (w/o Attn.)	87	0.316	0.222	0.365	0.5615
Our model (w Attn.)	**85**	**0.335**	**0.243**	**0.383**	**0.59**
FB15k-237					
DKRL	90.5	0.298	0.187	0.337	0.523
Our model (w/o Attn.)	90.5	0.314	0.217	0.349	0.527
Our model (w Attn.)	**90**	**0.316**	**0.229**	**0.356**	**0.545**

REFERENCES

[1] Russa Biswas, Mehwish Alam, and Harald Sack, 'Is aligning embedding spaces a challenging task? an analysis of the existing methods', *arXiv preprint arXiv:2002.09247*, (2020).

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko, 'Translating Embeddings for Modeling Multi-Relational Data', in *NIPS*, (2013).

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, 'Word translation without parallel data', *arXiv preprint arXiv:1710.04087*, (2017).

[4] Yuanfei Dai, Shiping Wang, Neal N Xiong, and Wenzhong Guo, 'A survey on knowledge graph embedding: Approaches, applications and benchmarks', *Electronics*, (2020).

[5] Tim Dettmers, Minervini Pasquale, Stenetorp Pontus, and Sebastian Riedel, 'Convolutional 2d knowledge graph embeddings (2018)', in *AAAI*.

[6] Jun Feng, Minlie Huang, Yang Yang, and Xiaoyan Zhu, 'GAKE: Graph aware knowledge embedding', in *COLING*, (2016).

[7] Genet Asefa Gesese, Russa Biswas, Mehwish Alam, and Harald Sack, 'A survey on knowledge graph embeddings with literals: Which model links better literal-ly?', *arXiv preprint arXiv:1910.12507*, (2019).

[8] Giuseppe Pirrò, 'Explaining and suggesting relatedness in knowledge graphs', in *ISWC 2015*.

[9] Nils Reimers and Iryna Gurevych, 'Sentence-bert: Sentence embeddings using siamese bert-networks', in *EMNLP-IJCNLP*.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, 'Sequence to sequence learning with neural networks', in *NIPS*, (2014).

[11] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun, 'Representation Learning of Knowledge Graphs with Entity Descriptions', in *AAAI*, (2016).