University of Fribourg / Faculty of Science and Medicine / Department of Informatics

Extending Knowledge Graph Embeddings for Data Imputation

Paolo Rosso

With the advancement of Big Data and Natural Language Processing (NLP) technologies, extensive research into Knowledge Graphs (KGs) has been conducted. In a typical KG, such as Wikidata, entities are connected via relations. A popular approach to represent facts in KGs is to define them as triplets (head, relation, tail). For example, the fact Bern, capitalOf, Switzerland, is composed of two entities, Bern and Switzerland, connected by the relation capitalOf. Although KGs are effective in representing structured data, they cannot be used to train modern Machine Learning models which often require numerical input. To tackle this issue, Knowledge Graph embeddings have been proposed. In our context, KG embeddings aim to project entities and relations from a KG into a low-dimensional and continuous vector space. The main benefit of such a representation is that the resulting vectors can be subsequently used as input to Machine Learning pipelines. In this thesis, we first introduce popular Knowledge Graphs, as well as typical KG embedding models. After providing an overview of the applications and problems that can be tackled with KG embeddings, we present our own contributions to this research field. Specifically, we first propose a novel approach, called JOINER, to jointly learn KG embeddings from text and a Knowledge Graph by taking advantage of both large-scale unstructured content (text) and high-quality structured data (the Knowledge Graph). JOINER not only preserves co-occurrences between words in a text corpus and relations between entities in a Knowledge Graph, it also provides the flexibility to control the amount of information shared between the two data sources via regularization. We conduct a thorough evaluation of JOINER on three evaluation tasks (analogical reasoning, link prediction and relation extraction) using three different corpora, showing significant improvement on most tasks. Next, we present a new KG embeddings model, called HINGE, able to learn hyper-relational facts from KGs, which are facts containing not only a base triplet (head, relation, tail) but also associated key-value pairs. HINGE captures not only the primary structural information of the KG encoded in the triplets, but also the correlation between each triplet and its associated key-value pairs. Our extensive evaluation shows the superiority of HINGE on various link prediction tasks over KGs, outperforming not only the KG embedding methods learning from triplets only (by 0.81-41.45%), but also the methods learning from hyper-relational facts using an n-ary representation (by 13.2-84.1%). Additionally, we propose an end-to-end solution called RETA in order to tackle instance completion problems by suggesting relation-tail pairs given a head entity. RETA consists of two components: RETA-Filter and RETA-Grader. More precisely, RETA-Filter first generates a filtered list of candidates by extracting and leveraging the schema of a KG; RETA-Grader then evaluates and ranks the candidate pairs considering the plausibility of both the candidate triplet and its corresponding schema using a newly designed KG embedding model. We evaluate our methods against a sizable collection of state-of-the-art techniques on three real-world KG datasets. Results show that our RETA-Filter generates of high-quality candidate r-t pairs, outperforming the best baseline techniques while reducing by 10.61%-84.75% the candidate pool size under the same candidate quality guarantees. Moreover, our RETA-Grader also significantly outperforms state-of-the-art link prediction techniques on the instance completion task by 16.25%-65.92% across different datasets. Finally, we address research questions raised in this thesis related to a number of Knowledge Graph embedding methods presented in the next sections. Additionally, we summarize the series of contributions we made in some of the core tasks tackled by Knowledge Graph embedding methods and conclude our thesis by discussing how to extend our proposed works.

Jury:
Prof. Dr. Denis Lalanne, University of Fribourg (Switzerland), president of the jury.
Prof. Dr. Philippe Cudré-Mauroux, University of Fribourg (Switzerland), thesis supervisor.
Prof. Dr. Dingqi Yang, University of Macau (Macau), thesis co-supervisor.
Dr. Bryan Perozzi, Google (NYC, United States of America), external examiner.
Prof. Dr. Jie Tang, Tsinghua University (Beijing, China), external examiner.