Summary of the Thesis Titled: "Neural Networks for Building Semantic Models and Knowledge Graphs"

Giuseppe Futia

2020

Knowledge Graphs (KGs) have emerged as a core abstraction for incorporating human knowledge into intelligent systems. This knowledge is encoded in a graph-based structure, whose nodes represent real-world entities, while edges define meaningful and binary relations between these entities. KGs are gaining attention from both the industry and the academia, because they provide a flexible way to capture, organize, and explore large amount of multi-relational data by means of queries. The structured knowledge that shapes KGs can be composed of simple statements, such as "Socrates is a man", or quantified statements, such as "All men are mortal". Simple statements represent a collection of *facts* that are arranged as edges within the KGs, while quantified statements require a more advanced and expressive form to represent knowledge. Ontologies define a standard formalism that enable this meaningful representation, specifying the semantics of entities and relations adopted to label nodes and edges of KGs.

Deductive and inductive reasoning approaches allow to extend a KG, improving its underlying and structured knowledge. On the one hand, deductive methods employ the simple statements and the set of rules defined by ontologies to derive additional knowledge, for instance 'Socrates is mortal", which is logically interpretable and understandable. On the other hand, inductive techniques involve simple and quantified statements to create further knowledge, discovering and generalising patterns available in the KG. These patterns can be inferred by the application of statistical learning methods on multi-relational data, which are less interpretable than deductive approaches, but they are capable to exploit latent factors in the KG for specific purposes, e.g. link prediction. The most recent implementation of these statistical learning methods include representation learning techniques, based on deep architectures of Neural Networks (NNs). These architectures contributed to reaching unprecedented results in prediction and classification tasks of modern Artificial Intelligence (AI) systems. More specifically, at the time of writing this thesis, NNs natively-built

for graph structures, the so-called Graph Neural Networks (GNNs), are gaining momentum and empowering the cutting-edge research on graph data.

The main goal of this thesis is to investigate the potential role of NN architectures, in particular the GNNs, to address two main open problems in the KGs research field: (i) the automatic building of semantic models of data sources, which represents a key factor for publishing reach semantic content into large-scale KGs; (ii) the automatic refinement of existing KGs by inferencing soft, but consistent knowledge in terms of new edges (or links). Such new edges are hard to encode into deductive and logic-based reasoning, but are extremely useful to develop tools on top of KGs, e.g. recommendation systems. Furthermore, this thesis reports the results within two different application domains, the public procurement and the academic publications, to show the impact of NNs in real scenarios.

In regards to the first open problem, this thesis illustrates novel contributions to the automatic semantic modeling with NN architectures. An initial study is conducted applying a simple, but efficient neural language model, such as Word2Vec, on SPARQL queries performed on different KGs. However, this approach does not take full advantage of the graph structure for the learning process: SPARQL queries include a limited number of graph patterns and Word2Vec treats these patterns as plain text. Considering these limits, a deeper investigation is conducted, developing a tool called SeMi (SEmantic Modeling machIne), which employs a novel method based on Graph Neural Networks (GNNs), trained on available multi-relational data repositories. The goal of this approach is to produce a latent representation of the entities and the relations between these entities, from the graph structure of the multi-relation data adopted as training set. The results of this investigation show that the adoption of these latent representations increases the accuracy of the computed semantic models, compared to manually-selected features. SeMi has been adopted in a real scenario to support the building of a novel KG in the public procurement domain.

In regards to the second open problem, the thesis reports an approach based on GNNs, for predicting new edges within a novel KG developed in the field of scholarly data. In this context, the thesis presents Geranium, a semantic platform to collect and organize the scientific knowledge of the Politecnico di Torino (Polito). The research achievements obtained with Geranium are the following: (i) an academic KG that semantically connects information on researchers and publications of Polito; (ii) a semantic search engine that aggregates such information and enables advanced features for the content exploration; (iii) a recommendation system which exploits the above-mentioned link prediction mechanism to suggest, for instance, novel collaboration opportunities between researchers of different disciplines, who worked on the same topics.

This thesis shows auspicious results in the adoption of inductive approaches based on NNs to address the above-mentioned research problems in the field of KGs. However, alongside this encouraging progress, inductive techniques do not provide human-understandable insights on how a specific result was achieved. Furthermore, the applications domains such as those analyzed in this thesis —

public procurement and academic publications — are contexts where the impact of NNs is relevant: the interpretability of the results is not only a desirable property, but it is a fundamental requirement for the stakeholders involved in these domains. Nevertheless, most of the available approaches to implement an eXplainable Artificial Intelligence (XAI) focus on technical solutions usable only by experts able to understand and manipulate the computational architectures of NNs. A complementary approach could incorporate deductive methods, which are able to exploit the symbolic representation of KG for inference new logic-based knowledge. The final part of this thesis present new research trajectories in the field, proposing neural-symbolic integration as a cornerstone to design an AI which is closer to non-insiders comprehension. Within such a general direction, the thesis proposes three specific challenges for future research—knowledge matching, cross-disciplinary explanations and interactive explanations.