Deep Learning on Academic Knowledge Graphs

Supervisors
Dr. Antonio Vetrò
Prof. Juan Carlos De Martin

Candidate
Giovanni Garifo

16 December 2019

The publication and sharing of research results is one of the main goals of academic institutions. In recent years, many efforts have been made to collect and organize the scientific knowledge through new, more accessible and comprehensive data repositories.

An example of such tools is the Institutional Research Information System (IRIS), an institutional publication repository developed by the Cineca Consortium and used by the Politecnico di Torino to store and share all the scientific papers published by its researchers. IRIS allows to explore the papers by searching within a field of study, matching the search terms with the keywords used by the authors to tag their publications.

The current implementation has some limitations: being inserted by the authors, the keywords can be acronyms, can contain abbreviations, initials written without capital letters, or misspelled words. In addition, they do not represent unique and field-wide semantic concepts, but they are simple character strings. As a consequence, the search engine of IRIS is incapable of correctly retrieve all the publications about a specific research topic, being unable to match the searched field of study with an unambiguous semantic entity.

The goal of this work is to experiment a new approach based on semantic technologies and deep neural networks in order to address the abovementioned limitations and enabling new opportunities to explore insights about the scientific community of the Politecnico di Torino.

The main contributions are:

1. A novel Knowledge Graph (KG) built on top of the scholarly data produced by researchers at Politecnico di Torino.
2. A deep learning algorithm for the prediction of new facts in this KG.
3. A recommendation system for the suggestion of useful insights based on such new facts.

Concerning the first contribution, Knowledge Graphs are a particular class of graphs that are used to semantically describe the human knowledge in a specific domain by linking semantic entities through labeled and directed edges. In the latest years many private and public organizations used KGs to organize and store data in a semantically coherent way. An example is the Google Knowledge Graph, which is used to enhance the Google search engine and virtual assistant capabilities, or the Open Academic Graph, a scientific Knowledge Graph made by Microsoft and AMiner that describes more than three hundred million academic papers, and is used to study citation networks and papers content.

To build the academic graph, a dump of the IRIS database has been used as input data. The dump contains more than 20,000 papers, each including relevant metadata like title, authors, abstract, date, type and venue of publication. Each publication is linked to semantic topics by means of TellMeFirst, a tool previously developed at the Nexa Center for Internet & Society. The semantic topics are added to the Knowledge Graph as uniquely identified entities, thus solving the problems related to the ambiguity of the keywords inserted by the publications authors. In addition, the resulting graph links together publications, researchers, semantic topics and scientific journals. This new academic graph has been called the Polito Knowledge Graph (PKG).

The availability of the PKG is the enabler for the second contribution, i.e. a deep learning algorithm capable of taking as input graph data and use it for predicting non-existent links: this is one of the most challenging tasks in the field of statistical relational learning for graph data, because obtaining meaningful predictions is strictly dependent on the ability of the trained model to embed the graph nodes characteristics.

Relational Graph Convolutional Network (R-GCN) is a recent deep learning model that proved to be well suited to work with highly irregular structures such as graphs, and it has been used in this work to empower a link

predictor able to provide predictions about new facts in the knowledge base. Once such predictions have been obtained, they were translated into statements and added to the Polito Knowledge Graph. A manual and sample-based validation confirmed that most of the predicted facts are meaningful.

Finally, the third contribution is a visualization of the predicted facts through a recommendation system for the suggestion of useful insights such as topics matching with researchers interests or journals accepting publications in their research field. The recommendation system is a first step in the direction of offering new tools to the researchers for exploring both new research fields or discovering researchers from other disciplines working in the same field.

I conclude remarking that the Polito Knowledge Graph and its recommendation system can be also used by administration offices of the Polito di Torino and from external parties – such as private organizations or government agencies – who may be interested in finding expert profiles in a given research area, breaking the current silos of keyword-based searches or discipline-specific knowledge bases.