Deep Learning on Academic Knowledge Graphs

Supervisors
Prof. Antonio Vetrò
Prof. Juan Carlos De Martin

Candidate
Giovanni Garifo

16 December 2019

The publication and sharing of new research results is one of the main goal of an academic institution. 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 IRIS, which stores all the scientific papers published by the Politecnico di Torino researchers. IRIS allow to explore the published papers by searching for a field of study, matching it with some keywords that are inserted by the authors and used to tag the publication. Such keywords do not represents unique and system-wide semantic concepts, but are simple character strings that are matched with the user input.

The current implementation has some limitations: being inserted by the authors, such keywords can be acronyms, can contain abbreviations, initials written without capital letters, or misspelled words. The consequence is that the search engine of IRIS is unable to correctly retrieve all the publications about a specific research topic, because the system cannot match the searched field of study with an unambiguous *semantic entity*, but only with character strings that are not uniquely identified or semantically linked each other, and also prone to lexical errors.

The goal of this work is to overcome such limitations, enabling new possibilities for exploring and obtaining insights about the scientific community of the Politecnico di Torino through the use of a new semantic-empowered search engine and recommendation system. However, at the foundations of this tool there must be a new data structure capable of representing semantic relations and concepts.

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 have used such new kind of data structure 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 collects more then three hundred million academic papers, which is used to study citation networks and papers content.

In this work is presented a novel semantic graph built on top of the scholarly data produced by the Politecnico di Torino researchers. It is also presented how state-of-the-art machine learning algorithms have been used for the prediction of new facts in the knowledge base, and how such predicted facts are used to empower a recommendation system.

To build such academic graph, a dump of the IRIS database has been used as input data. The dump contains more then 20,000 papers, together with their relevant metadata, like the title, the authors, the abstract, the journal and the date of publication. In order to link each publication to its research topics we employed TellMeFirst, a tool for the automatic extraction of semantic concepts from texts, which uses DBpedia as its source of knowledge. Using the publications abstracts as input text for TellMeFirst, we have been able to extract such semantic topic and add them in the knowledge graph as uniquely identified entities, thus solving the problems related to the ambiguity of the keywords inserted by the publications authors.

The resulting graph links together publications, researchers, fields of study and scientific journals in order to build a knowledge base that describes the Politecnico di Torino scientific community. This new academic graph has been called the *Polito Knowledge Graph*.

The availability of such a complex and informative data structure leads to the opening of interesting scenarios, especially when thinking about the latent information that can be extracted from it. In recent years, efforts have been made to develop machine learning algorithms capable of taking as input graph data, both for the classification of unseen nodes and for the prediction of non-existent links. The latter is one of most challenging tasks in the field of statistical relational learning for graph data, mainly because in order to obtain meaningful predictions it is mandatory to learn models that are able to truly embed in their parameters the characteristics of the graph nodes.

Deep Learning techniques have recently proved to be well suited for the developement of embedding models. In fact, new architectures derived from the image recognition field have been specifically built to work with highly irregular structures, such as graphs. In particular, the Relational Graph Convolutional Network (R-GCN) has been specifically developed to work with relational graph. Such new kind of neural network can be used to build

encoding models that are able to obtain a node vector representation by convolving some self-learnt filters with the features of its adjacent nodes.

In order to predict new facts inside the Polito Knowledge Graph, firstly a useful dataset for the learning task has been built starting from its internal representation. Then, such dataset has been used to train an encoder model (based on a R-GCN) which is capable of obtaining nodes embeddings directly learnt from the graph structure itself, without requiring any prior knowledge or feature engineering. Such learnt embeddings are then used to obtain meaningful predictions about new facts in the knowledge base by means of a decoder model. The decoder creates the set of every possible edges in the graph and scores every edge by means of a scoring function. Such function takes as input the vector representations of the source and destination nodes of an edge, and uses a factorization methods to obtain the likelihood for such edge to represent a true fact of the knowledge base.

Once such predictions are obtained, they can be translated into statements and added to the Polito Knowledge Graph, thus completing its knowledge base.

Looking at the predictions obtained, it is possible to see how the model has been able to truly characterize the graph nodes, thus predicting meaningful facts. For example, the model predicted for a publication regarding real-time tools for emergency management a new edge that links such publication to the entity regarding the list of software reliability models, which looks like an appropriate suggestion. The model also predicted a new edge that connects the above publication to a researcher whose main field of research is in integrated circuits for aircraft applications, which is again a good prediction being the two topics fairly related. The above examples are just two among all the new, meaningful facts that the model has been able to predict.

The predicted facts can then be showed by a recommendation system for the suggestion of useful insights to the Politecnico di Torino researchers. Such recommendation system is a first step in the direction of offering new tools to the researchers for exploring both new research fields and the scientific community that study such fields.

Moreover, the Polito Knowledge Graph is used to empower an expertise search engine made available to both the administration offices of the Polito di Torino and to external parties, such as private organizations or government agencies, who may be interested in finding expert profiles in a given research area.