Special Issue on Machine Learning and Knowledge Graphs

Mehwish Alam[a], Anna Fensel[b,1], Jorge Martinez-Gil[d], Bernhard Moser[d], Diego Reforgiato Recupero[e], Harald Sack[a]

[a] *FIZ Karlsruhe, Leibniz Institute for Information Infrastructure, Karlsruhe, Germany*
[b] *Wageningen Data Competence Center, Consumption and Healthy Lifestyles Chair Group, Wageningen University and Research, Wageningen, the Netherlands*
[c] *Semantic Technology Institute (STI) Innsbruck, Department of Computer Science, University of Innsbruck, Innsbruck, Austria*
[d] *Software Competence Center Hagenberg, Hagenberg, Austria*
[e] *University of Cagliari, Cagliari, Italy*

Preface

Machine Learning and Knowledge Graphs (KGs) are currently essential technologies for designing and building large scale distributed intelligent systems. Machine Learning is a well established field, which has currently gained a high momentum due to the advances in the computational infrastructures, availability of Big Data, and appearance of new algorithms based on Deep Learning. In fact, Deep Learning methods have become an important area of research, achieving some important breakthrough in various research fields, especially Natural Language Processing (NLP) as well as Image and Speech Recognition.

KGs are large networks of real-world entities described in terms of their semantic types and their relationships to each other. Examples include the Google Knowledge Graph[1] with over 70 billion facts (in 2016), dataCommons[2], DBpe-

Email addresses: mehwish.alam@fiz-karlsruhe.de (Mehwish Alam), anna.fensel@wur.nl (Anna Fensel), Jorge.Martinez-Gil@scch.at (Jorge Martinez-Gil), Bernhard.Moser@scch.at (Bernhard Moser), diego.reforgiato@unica.it (Diego Reforgiato Recupero), harald.sack@fiz-karlsruhe.de (Harald Sack)

[1] https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

[2] https://datacommons.org/

dia [1][3], YAGO [2] and YAGO2 [3][4], Wikidata[5] and Knowledge Vault [4], a very large scale probabilistic knowledge graph created with information extraction methods for unstructured or semi-structured information. Specifically, KGs provide the means of development of the newest methods for data management, data fusion / data merging, graph and network optimization and modeling, serving as a source of high quality data and a base for ubiquitous information integration. While KGs provide explicit knowledge representations in terms of underlying ontologies based on symbolic logic, Machine Learning more specifically deep learning algorithms which provide implicit or latent representations of the knowledge graphs.

Although Machine Learning (and Deep Learning) and KG technologies have been deployed separately, in the last years, the first works combining these technologies are showing large potential in solving many real-world challenges such as Scholarly data [5], Knowledge Reconciliation [6], etc. This is largely due to recent work on KG embeddings that allows for the projection of KGs to low dimensional vector spaces. These methods can further be used for KG completion tasks such as head and tail prediction, relation prediction and triple classification [7].

One starting point for a look into the future is the analysis of the strengths and shortcomings of the respective disciplines. The success of various Machine Learning methods, in particular Deep Neural Networks (DNNs), for challenging problems of computer vision and pattern recognition, has led to a Cambrian explosion in the field of Artificial Intelligence (AI).

In many application areas, AI researchers have turned to Deep Learning as the solution of choice [8, 9]. A characteristic of this development is the acceleration of progress in AI over the last decade, which has led to AI systems that are strong enough to raise serious ethical and societal acceptance questions.

[3] `https://wiki.dbpedia.org/`

[4] `https://yago-knowledge.org/`

[5] `https://www.wikidata.org/wiki/Wikidata:Main_Page`

It is an amazing, unexpected phenomenon that deep networks become easier to be optimized (trained), that is to find a reasonable sub-optimum out of many equally good possibilities, with an increasing number of nodes and layers, hence complexity, see [10, 11]. But, the increased size of dimensionality of the models' configuration space brings about inherent problems. For example, there is the still unsolved problem of lack of control of high-dimensionality effects [12] which can cause instabilities as illustrated, for example, by the emergence of so-called adversarial examples [13, 14].

In contrast to traditional engineering, there is a lack of uniqueness of internal configuration, causing difficulties in model comparison. Therefore, in particular Deep Learning models, systems based on Machine Learning are typically regarded as black boxes affecting the correct interpretation of the system's output and the transparency of the system's configuration. The problem is challenging, and it is not just simply the complex nested non-linear structure that matters, as often pointed out in the literature, see [15]. There are mathematical or physical systems that are also complex, nested, and non-linear, yet stable, interpretable, and controllable (e.g., wavelets, statistical mechanics). Instead, the problem is related to a lack of structure and constraints, whether in an algebraic or logical-semantic sense. However, complex Deep Learning models are based on evaluating samples and assumptions about the model space. Hence, the more samples, the better the expected accuracy.

In contrast to today's Machine Learning, human intelligence distinguishes its capability to take context into account. One aspect of context refers to links to conditions that are not explicitly represented in the training data. Context data are not simply additional data, such as a different sample drawn from an unknown distribution, but context data provide other meta-information that helps impose restrictions on the interpretation and use of the observed samples and avoid potential misinterpretations. In contrast to a purely statistical approach as it is found today in Machine Learning, knowledge graphs offer an interesting perspective due to their flexibility in representing linked context data from different sources and formats. Hence, the more linked context data, the bet-

ter the expected robustness. A future perspective, therefore, lies in combining both approaches, statistical learning, and contextual reasoning. The emerging discipline of Relational Machine Learning addresses this endeavor [16, 17].

Further, we describe the aims and contents of our special issue.

1. Aims

This Special Issue[6] brings together the achievements and the experiences of researchers and practitioners from several areas, including Machine Learning, Deep Learning, and Knowledge Graphs.

To pursue more advanced methodologies, it has become critical that the communities related to Machine Learning, Deep Learning, and Knowledge Graphs join their forces to develop more effective algorithms and applications. In particular, two main technology directions solicited for this special issue were as follows:

1. Improved Machine Learning with Knowledge Graphs: employing semantic models and linked data for the training steps, learning effective representation from the Knowledge Graphs for the tasks of feature extraction, classification, prediction and decision making. A successful example here includes IBM Watson question answering system, that has outperformed the best human players of an intellectual TV quiz show "jeopardy!"
2. Machine Learning for Improving Knowledge Graphs: while capturing semantics correctly is impossible without at least some human involvement, Machine Learning can assist the knowledge acquisition of the semantic structures substantially. For example, knowledge graphs can be created by employing Deep learning, and then subsequently verified by the humans. The aim is to gather the state-of-art research on adaptation, control, decision making and knowledge management in smart systems. A

[6] https://www.sciencedirect.com/journal/future-generation-computer-systems/special-issue/104N2HJH560

special focus is on the technological solutions that exploit agile, adaptive and reconfiguration characteristics to address unexpected circumstances and evolving scenarios that arise in smart systems.

We have been soliciting high-quality manuscripts reporting relevant research in the area of generation of knowledge graphs by using deep learning techniques. Topics of interest include, but are not limited to:

- Architectures for systems based on Machine Learning and Knowledge Graphs,

- Machine Learning and Knowledge Graphs in distributed large scale systems,

- Information management with on Machine Learning and Knowledge Graphs,

- Probabilistic Knowledge Graphs,

- Creation and maintenance (curation, quality assurance...) of Knowledge Graphs employing Machine Learning,

- Machine Learning techniques employing Knowledge Graphs,

- Explainable Artificial Intelligence for data-intensive systems based on Knowledge Graphs,

- Non-functional application of Knowledge Graphs in data-intensive systems e.g. for legal use of data, user consent solicitation, smart contracting, GDPR,

- New approaches for combining Deep Learning and Knowledge Graphs,

- Methods for generating Knowledge Graph (node) embeddings,

- Scalability issues,

- Temporal Knowledge Graph Embeddings,

- Applications of combining Deep Learning and Knowledge Graphs,

- Recommender Systems leveraging Knowledge Graphs,

- Link Prediction and completing KGs,

- Ontology Learning and Matching exploiting Knowledge Graph-Based Embeddings,

- Knowledge Graph-Based Sentiment Analysis,

- Natural Language Understanding/Machine Reading,

- Question Answering exploiting Knowledge Graphs and Deep Learning,

- Entity Linking,

- Trend Prediction based on Knowledge Graphs Embeddings,

- Domain Specific Knowledge Graphs (e.g. Smart Cities, Scholarly, Biomedical, Musical),

- Applying knowledge graph embeddings to real world scenarios.

2. Contents

The special issue on Machine Learning and Knowledge Graphs attracted 15 submissions covering both Machine Learning and Knowledge Graphs technologies. Each paper was reviewed by at least three reviewers. The papers that have been accepted in this special issue covers several different domains and tasks. In each of them, Machine Learning techniques and Knowledge Graphs have been coupled for a particular purpose.

The first paper, "FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with Machine Learning", by Bram Steenwinckel, Dieter De Paepe, Sander Vanden Hautte, Pieter Heyvaert, Mohamed Bentefrit, Pieter Moens, Anastasia Dimou, Bruno Van Den Bossche, Filip De Turck, Sofie Van Hoecke and Femke Ongenae, [18] uses semantic knowledge to improve a mixture of data-driven and knowledge-driven techniques to detect anomalies and faults. Moreover, semantic rule mining methods are adopted to increase adaptiveness through the feedback of faults and anomalies.

The second paper, "Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain", by Danilo Dessi, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi and Enrico Motta, [19] tackles the challenge of knowledge extraction by employing several state-of-the-art Natural Language Processing and Text Mining tools. Moreover, it describes an approach for integrating the generated entities and relationships. The authors generated a scientific knowledge graph

including 109,105 triples, extracted from 26,827 abstracts of papers within the Semantic Web domain.

The third paper, "Linking OpenStreetMap with knowledge graphs — Link discovery for schema-agnostic volunteered geographic information", by Nicolas Tempelmeier and Elena Demidova, [20] proposes a novel link discovery approach to predict identity links between Open Street Map nodes and geographic entities in a knowledge graph. The core of the approach is a novel latent, compact representation of Open Street Map nodes that captures semantic node similarity in an embedding. This latent representation is used to train a supervised model for link prediction and utilises existing links between Open Street Map and knowledge graphs for training.

The fourth paper, "Improving University Faculty Evaluations via multi-view Knowledge Graph", by Qika Lin, Yifan Zhu, Hao Lu, Kaize Shi and Zhendong Niu, [21] proposes a novel University Faculty Evaluation System based on a multi-view Knowledge Graph that integrates heterogeneous faculty data. By integrating the academic development status of scholars in the previous three years as well as student evaluation data, this paper proposes an academic development factor for making predictions about faculty academic development whose experiments show that this factor is closely related to the features of the knowledge graph and student evaluations.

The fifth paper entitled "Neural machine translating from natural language to SPARQL", by Xiaoyu Yin, Dagmar Gromann and Sebastian Rudolph, [22] evaluates the utilization of eight different Neural Machine Translation models for the task of translating from natural language to the structured query language SPARQL. The results show a dominance of a Convolutional Neural Network based approach.

The sixth paper, "Extracting knowledge from Deep Neural Networks through graph analysis", by Vitor A.C. Horta, Ilaria Tiddi, Suzanne Little and Alessandra Mileo, [23] introduces the concept of a co-activation graph and investigates the potential of graph analysis for explaining deep representations. The co-activation graph encodes statistical correlations between neurons' activation

values and therefore helps to characterise the relationship between pairs of neurons in the hidden layers and output classes. The findings of the paper show that graph analysis can reveal important insights into how Deep Neural Networks work and enable partial explainability of Deep Learning models.

The seventh paper, "On the impact of knowledge-based linguistic annotations in the quality of scientific embeddings", by Andres Garcia-Silva, Ronald Denaux and Jose Manuel Gomez-Perez, [24] includes a comprehensive study on the use of explicit linguistic annotations to generate embeddings from a scientific corpus and on the assessment of their impact in the resulting representations. Results show how the effect of such annotations in the embeddings varies depending on the evaluation task. In general, it turns out that learning embeddings using linguistic annotations contributes to achieve better evaluation results.

The eighth paper, "Instance Matching in Knowledge Graphs through random walks and semantics", by Ali Assi and Wajdi Dhifli, [25] proposes a novel approach for the Instance Matching problem based on Markov random walks. The proposed approach leverages both the local and global information mutually calculated from a pairwise similarity graph. Semantic and bipartite graph-based post-processing strategies that operate on the obtained random walk ranks to optimize the final assignment of co-referents are proposed. A scalable distributed implementation of the proposed approach on top of the Spark framework is also implemented and evaluated on benchmark datasets from the instance track of the Ontology Alignment Evaluation Initiative.

The ninth paper, "Representing emotions with knowledge graphs for movie recommendations", by Arno Breitfuss, Karen Errou, Anelia Kurteva and Anna Fensel, [26] proposes a knowledge graph representing human emotions within the domain of movies. It is based on sentiment analysis [27, 28] and it was constructed by retrieving emotions from movie reviews through Machine Learning approaches. A chatbot using a reasoning mechanism combined with users' emotions has been built to recommend movies. It turned out that it is feasible but it needs more information about the emotions associated with the movies

than those publicly available online.

The tenth paper, "Continual representation learning for node classification in power-law graphs", by Gianfranco Lombardo, Agostino Poggi, Michele Tomaiuolo, [29]. This paper proposes an approach for learning representations on graphs where the nodes are evolving over time and it is needed that the algorithm learns the representations of only the new nodes. The method is based on a continual feature learning meta-algorithm for node embedding. The authors show that the light weight solution based on hub nodes for encoding new nodes perform better or comparable on the task of node labeling as compared to static approaches.

Acknowledgements

This research was co-funded by Interreg Austria-Bavaria project KINet (grant agreement: AB 292).

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, Vol. 6, 2007, pp. 722 735. doi:10.1007/978-3-540-76298-0_52.

[2] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: WWW, 2007.

[3] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum, Yago2: A spatially and temporally enhanced knowledge base from wikipedia, Artificial Intelligence 194 (2013) 28 – 61, artificial Intelligence, Wikipedia and Semi-Structured Resources. doi:https://doi.org/10.1016/j.artint.2012.06.001.
URL http://www.sciencedirect.com/science/article/pii/S0004370212000719

[4] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zhang, Knowledge vault: A web-scale approach to probabilistic knowledge fusion, in: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, 2014, pp. 601 610, evgeniy Gabrilovich Wilko Horn Ni Lao Kevin Murphy Thomas Strohmann Shaohua Sun Wei Zhang Geremy Heitz.
URL http://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf

[5] M. Nayyeri, G. M. Cil, S. Vahdati, F. Osborne, A. Kravchenko, S. Angioni, A. A. Salatino, D. R. Recupero, E. Motta, J. Lehmann, Link prediction of weighted triples for knowledge graph completion within the scholarly domain, IEEE Access 9 (2021) 116002–116014. doi:10.1109/ACCESS.2021.3105183.
URL https://doi.org/10.1109/ACCESS.2021.3105183

[6] M. Alam, D. R. Recupero, M. Mongiovì, A. Gangemi, P. Ristoski, Event-based knowledge reconciliation using frame embeddings and frame similarity, Knowl. Based Syst. 135 (2017) 192–203. doi:10.1016/j.knosys.2017.08.014.
URL https://doi.org/10.1016/j.knosys.2017.08.014

[7] G. A. Gesese, R. Biswas, M. Alam, H. Sack, A survey on knowledge graph embeddings with literals: Which model links better literal-ly?, Semantic Web 12 (4) (2021) 617 647. doi:10.3233/SW-200404.
URL https://doi.org/10.3233/SW-200404

[8] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (FEB) (2015) 436 444.

[9] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, P. Corke, The limits and potentials of deep learning for robotics, The International Journal of Robotics Research 37 (4-5) (2018) 405–420.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, International Conference on Learning Representations.

[11] R. Vidal, J. Bruna, R. Giryes, S. Soatto, Mathematics of deep learning, arXiv e-prints.

[12] A. N. Gorban, I. Y. Tyukin, Blessing of dimensionality: mathematical foundations of the statistical physics of data, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376 (2118).

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv e-prints.

[14] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples, arXiv e-prints`arXiv:1707.07397`.

[15] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, arXiv e-prints.

[16] M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, A review of relational machine learning for knowledge graphs, Proceedings of the IEEE 104 (1) (2016) 11–33.

[17] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications., IEEE Trans. Knowl. Data Eng. 29 (12) (2017) 2724—2743.

[18] B. Steenwinckel, D. De Paepe, S. Vanden Hautte, P. Heyvaert, M. Bentefrit, P. Moens, A. Dimou, B. Van Den Bossche, F. De Turck, S. Van Hoecke, F. Ongenae, Flags: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning, Future Generation Computer Systems 116 (2021)

30–48. doi:https://doi.org/10.1016/j.future.2020.10.015.
URL https://www.sciencedirect.com/science/article/pii/S0167739X20329927

[19] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain, Future Generation Computer Systems 116 (2021) 253–264. doi:https://doi.org/10.1016/j.future.2020.10.026.
URL https://www.sciencedirect.com/science/article/pii/S0167739X2033003X

[20] N. Tempelmeier, E. Demidova, Linking openstreetmap with knowledge graphs — link discovery for schema-agnostic volunteered geographic information, Future Generation Computer Systems 116 (2021) 349–364. doi:https://doi.org/10.1016/j.future.2020.11.003.
URL https://www.sciencedirect.com/science/article/pii/S0167739X20330272

[21] Q. Lin, Y. Zhu, H. Lu, K. Shi, Z. Niu, Improving university faculty evaluations via multi-view knowledge graph, Future Generation Computer Systems 117 (2021) 181 192. doi:https://doi.org/10.1016/j.future.2020.11.021.
URL https://www.sciencedirect.com/science/article/pii/S0167739X20330454

[22] X. Yin, D. Gromann, S. Rudolph, Neural machine translating from natural language to sparql, Future Generation Computer Systems 117 (2021) 510 519. doi:https://doi.org/10.1016/j.future.2020.12.013.
URL https://www.sciencedirect.com/science/article/pii/S0167739X20330752

[23] V. A. Horta, I. Tiddi, S. Little, A. Mileo, Extracting knowledge from deep neural networks through graph analysis, Future

Generation Computer Systems 120 (2021) 109–118. doi:https://doi.org/10.1016/j.future.2021.02.009.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21000613

[24] A. Garcia-Silva, R. Denaux, J. M. Gomez-Perez, On the impact of knowledge-based linguistic annotations in the quality of scientific embeddings, Future Generation Computer Systems 120 (2021) 26–35. doi:https://doi.org/10.1016/j.future.2021.02.019.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21000716

[25] A. Assi, W. Dhifli, Instance matching in knowledge graphs through random walks and semantics, Future Generation Computer Systems 123 (2021) 73–84. doi:https://doi.org/10.1016/j.future.2021.04.015.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21001369

[26] A. Breitfuss, K. Errou, A. Kurteva, A. Fensel, Representing emotions with knowledge graphs for movie recommendations, Future Generation Computer Systems 125 (2021) 715–725. doi:https://doi.org/10.1016/j.future.2021.06.001.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21001953

[27] A. Dridi, M. Atzeni, D. R. Recupero, Finenews: fine-grained semantic sentiment analysis on financial microblogs and news, Int. J. Machine Learning & Cybernetics 10 (8) (2019) 2199–2207. doi:10.1007/s13042-018-0805-x.
URL https://doi.org/10.1007/s13042-018-0805-x

[28] A. Dridi, D. R. Recupero, Leveraging semantics for sentiment polarity detection in social media, Int. J. Machine Learning & Cybernetics 10 (8) (2019) 2045 2055. doi:10.1007/s13042-017-0727-z.
URL https://doi.org/10.1007/s13042-017-0727-z

[29] G. Lombardo, A. Poggi, M. Tomaiuolo, Continual representation learning for node classification in power-law graphs, Future Generation Computer Systems 128 (2022) 420–428. doi:https://doi.org/10.1016/j.future.2021.10.011.
URL https://www.sciencedirect.com/science/article/pii/S0167739X21004015