Vol. 6 - N° 3 - September 2022

IJIST

International Journal of Information Science and Technology

Special Issue on Machine Learning and Natural Language Processing

GUEST EDITORS

Dr. Vito Pirrelli
Dr. Horacio Rodriguez
Dr. Arsalane Zarghili

PAPERS

ISSN 2550-5114

Machine Learning and Natural Language Processing. An editorial

Vito Pirrelli *vito.pirrelli@ilc.cnr.it*
Institute for Computational Linguistics
Italian National Research Council, Pisa, Italy

Horacio Rodriguez *horacio@lsi.upc.edu*
Universitat Politècnica de Catalunya, Barcelona, Spain

Arsalane Zarghili *arsalane.zarghili@usmba.ac.ma,*
Laboratoire Systèmes Intelligents et Applications
Faculté des sciences et Techniques de Fès
Université Sidi Mohamed Ben Abdellah

Editorial

It has been observed (Raoult, 2010) that the exponentially growing rate of technological innovation of the last 25 years, together with the virtually unbounded availability of digital data, has profoundly changed our research attitudes. Scientists have been moving from a methodological framework dominated by theory-internal hypotheses, to data-driven and curiosity-driven research. In the not too distant past, research strategies based on strong theoretical assumptions had often little chances of being refuted, due to the limited availability of the data used to support a hypothesis, the difficulty to replicate experimental evidence and the potential conflict of interests induced by theoretical biases. By contrast, nowadays research strategies based on data and technologies are more open to challenge received wisdom, and less biased by theoretical assumptions and prejudices.

Interdisciplinary convergence is another positive side-effect of this revolutionary change. The need to increase data sharing and establish a common ground for data analysis has promoted an increasing awareness of differences and similarities between approaches and research goals, together with an interest in circumventing disciplinary barriers. This has allowed researchers to address common problems in a synergistic way, and tackle issues of both practical and theoretical interests from a common perspective (Pirrelli et al., 2020).

Having said that, now that the methodological pendulum has definitely swung to a strong data-centred attitude in research, it would be mistaken to take sides for a form of radical epistemological relativism, and claim that theoretical frameworks play no role in any scientific dispute. In fact, even if technological innovations often produce evidence that theories are not ready to account for, it is also true that technologies more and more often produce evidence that technologies themselves cannot explain. This is the case of at least some aspects of the success story of the deep learning technology, which can replicate aspects of intelligent behaviour without helping gain any principled understanding of the reasons for their success (Poggio, 2012). In this connection, the need for model transparency and interpretability invoked by the advocates of Explainable Artificial Intelligence does not only respond to the moral obligation of verifying a model's adherence to shared ethical, social and legal values. It also reflects the fundamental concerns for a responsible science, accountable for measurable advances in our understanding of complex problems.

The present special issue is an edited collection of papers that set a good example in balancing a technology-driven curiosity for data and practical problems with a genuine methodological interest in better understanding the phenomena being modelled and the ways they should be modelled.

In their paper "Named Entity Recognition for Specific Domains: Take Advantage of Transfer Learning" Sunna Torge, Waldemar Hahn, Lalith Manjunath and René Jäkel offer a general assessment of the advantage of applying a classifier that was pre-trained on a named entity recognition task with knowledge from a

specific domain, to the same task in a different domain. Although the experiment was intended to address the practical and ubiquitous problem of circumventing the shortage of domain-specific training data, its implications are much wider, as they involve the general applicability of the machine learning paradigm known as transfer learning (Baxter, 1998). Albeit extremely encouraging (training is reduced by 10 times, with best prediction scores on domain-specific NER), the paper's results realistically acknowledge the limitations of current algorithms for transfer learning and provide useful hints for possible ways ahead in the field.

Recommendation systems are software tools and techniques that provide suggested items to a user. An item can designate any element that can be offered to a user such as a product, a service, media items or a collection of item information. A recommendation system helps users to make their choice in an area where they have little information, to sort and evaluate possible alternatives. In "Hybrid filtering and semantic sentiment analysis by deep learning for recommendation systems", Badiâa Dellal-Hedjazi and Zaia Alimazighi address the issue of augmenting the quality of recommended feedback by integrating different sources of user's and item's information that are usually exploited independently. A hybrid architecture combining deep learning technology with topic modelling of users' product reviews (tweets) via Latent Dirichlet Allocation appears to provide an interesting avenue for technology hybridization in the context of user-tailored information filtering.

In the paper "Deep Deterministic Policy Gradient based portfolio management system", Firdaous Khemlichi, Hiba Chougrad, Youness Idrissi Khamlichi, Abdessamad El Boushaki and Safae Elhaj Ben Ali address the problem of Portfolio management, i.e. the decision making process of continuously reallocating an amount of funds into a number of different financial investment products, with the aim to maximize the return while restraining the risk. In this domain, a deep variant of Reinforcement Learning appears to be the tool of the trade, as it solves the hard problem of correlating immediate actions (i.e. selling or holding any one of an array of securities and their derivatives) with the delayed outcomes they produce (the future increase of an initial investment). The deep dimension of RL consists in using a LSTM to approximate the function relating inputs (environment states and rewards) to outputs (actions), by iteratively adjusting weights over the network connections along gradients that promise increasing rewards.The model is shown to achieve a higher rate of return (14%) compared to baseline strategies like "Uniform Buy And hold" or "Best Stock".

Addresses are meaningful, highly structured linguistic units, which contain a number of named entities, such as a street name or a postal code. They are formatted according to a variety of different standards and conventions, heavily dependent on language, domain and geographic context, so as to raise a formidable challenge to standard parsing strategies (e.g. editing distance algorithms) aimed at tagging their informational content. In their contribution to the present issue, entitled "Multinational Address Parsing: A Zero-Shot Evaluation", Marouane Yassine, David Beauchemin, François Laviolette and Luc Lamontagne face this issue from a multi-lingual, semantic perspective. They develop a single deep neural architecture capable of parsing addresses from multiple countries, and explore the possibility of zero-shot transfer from some countries' addresses to others', and parsing incomplete addresses. To achieve this, they combine sub-word embeddings (a semantic representation of words as character n-gram vectors) with zero-shot transfer learning and attention mechanisms (Vaswani et al., 2017), showing they can obtain, among other things, also near state-of-the-art performance in a number of languages based on training data from a different language.

All in all, the volume bears witness to mature, explanation-aware approaches to technological progress in the field, geared towards the solution of practical problems, but also bound to face significant scientific challenges through technology hybridization and interdisciplinary contamination between Natural Language Processing and machine learning approaches.

The present collection stems from the $6th$ edition of the Morocco section of the IEEE CiSt Conference on "Machine Learning and Natural Language Processing", originally scheduled to be held in Agadir on December 12-18 2020, and then postponed to June 5-12 2021, due to the Covid-19 pandemic. We gratefully acknowledge the Congress General Chair, professor Mohammed El Mohajir, and the Scientific and Organizing Committees of the event, for their constant scientific and managerial support.

Pisa, May 2022

References

Jonathan Baxter. Theoretical models of learning to learn. In Sebastian Thrun and Lorien Pratt (eds.), *Learning to learn*, pp. 71–94. Springer, 1998.

Vito Pirrelli, Ingo Plag, and Wolfgang U. Dressler (eds.). *Word Knowledge and Word Usage: A Cross-disciplinary Guide to the Mental Lexicon*. Mouton de Gruyter Berlin, 2020.

Tomaso Poggio. The levels of understanding framework, revised. *Perception*, 41(9):1017–1023, 2012.

Didier Raoult. Technology-driven research will dominate hypothesis-driven research: the future of microbiology. *Future Microbiology*, 5(2):135–137, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.