Vol. 3 - N° 5 - September 2019

ijist

International Journal of Information Science and Technology

Special Issue on Machine Learning and Natural Language Processing
Guest Editors | Vito Pirrelli, Ouafae Nahli, Mohamed Chahhou

PAPERS

ISSN 2550-5114

Machine Language Learning and Natural Language Processing

Vito Pirrelli
Ouafae Nahli
Mohamed Chahhou

Bio sketches

Vito Pirrelli is Research Director at the CNR Institute for Computational Linguistics “Antonio Zampolli” in Pisa, where he runs the “Physiology of Communication” laboratory. Former Chair of “NetWordS”, the European Science Foundation Programme on Word Structure in the Languages of Europe, and co-editor in chief of the journal “Lingue e Linguaggio”, he is Principal Investigator of the Italian strategic project “Readlet: Reading to understand”, an ICT-driven, large-scale investigation of early grade children's reading strategies. His research interests include machine language learning, psycho-computational modeling of the mental lexicon, disorders of verbal communication and language acquisition, natural language processing, language teaching and artificial intelligence.

Ouafae Nahli

Nahli Ouafae is full-time Researcher at the CNR Institute for Computational Linguistics "Antonio Zampolli" in Pisa. Her research interests and contributions cover a number of aspects of both Classical and Standard Modern Arabic, including morpho-syntactic and lexico-semantic issues in Arabic Natural Language Processing, computational analysis of literary texts, and lexico-ontological modeling.

Mohamed Chahhou

Mohamed Chahhou has graduated from the "Ecole Polythechnique of Brussels" and obtained a PhD in Computer Science at the University Hassan I (Morocco). He is currently a Professor at the University Abdelmalek Essaadi in Tetouan (Morocco). Besides his research activities, he is very involved in Applied Machine Learning and has won many international Data Science competitions on the Kaggle platform.

Editorial note

Learning is key to any form of intelligent behavior. The dynamic ability to adapt an individual response to an open, ever changing environment, based on past events and their distribution in time and space, is what enables optimal adaptation to multiple conditions, as well as maximization of survival chances and opportunities for success.

Language, arguably the most efficient system of social communication available to humans, makes no exception to this general principle. Recent interdisciplinary research in the cognitive sciences has demonstrated that patterns of use strongly affect how language is acquired, is used, and changes (Beckner et al. 2006). Words, phrases and utterances are investigated as dynamic processes, emerging from interrelated patterns of sensory experience, communicative and social interaction and psychological and neurobiological

mechanisms (Elman 2009). According to this view, human lexical information is never stable, time-independent or context-independent. Its content is continuously updated and reshaped as a function of when, why and how often it is accessed and processed. Such flowing activation states are more reminiscent of the wave/particle duality in quantum physics (Libben 2016) or the inherently adaptive, self-organizing behavior of biological dynamic systems (Beckner et al. 2009) than ever thought before.

We believe that full investigation of language systems is likely to benefit from the use of basic concepts from the toolkit of complexity theory in biological networks, such as emergence, non-linearity and adaptive self-organization (Larsen-Freeman & Cameron 2008). In particular, the recent emphasis on a growing integration between Natural Language Processing (NLP) and Machine Learning technology has turned the research spotlight from the formal properties of language computations and representations on how language computations and representations can be developed from experience. Such a shift of emphasis has questioned the traditional view of human language as an abstract formal system, based on an algebraic symbolic calculus. Nonetheless, even if we assume (following received wisdom) that language processing is an algebraic calculus combining smaller units (single sounds or syllables) into larger units (entire utterances or speeches), the central question that must be addressed is how basic combinatorial units are acquired and put together from input evidence. In the end, we may ignore what rules consist of and what representations they manipulate, or even wonder whether rules and representations exist at all. Learning represents a fundamental level of meta-cognition whereby intelligent systems can successfully do things without knowing how and why they are successful in the first place (Poggio 2012).

For our present concerns, this has two important implications: a theoretical consequence and practical one. From a theoretical perspective, we believe that machine learning is bridging the gap between our understanding of how language is algorithmically used (level 2 of Marr's (1982) hierarchy) and how language algorithms are implemented in the brain (level 3 of Marr's hierarchy). At the same time, on a more practical and application-oriented footing, we must acknowledge that much of the extraordinary success of Artificial Intelligence in recent years can be understood in the light of an apparent paradox: learning lies at the heart of the capacity of an intelligent system to optimally carry out a real-life task with few general learning principles, massive trial-and-error training, and relatively shallow knowledge about the nature of the task itself.

It is thus not surprising that the successful integration of machine learning technology and NLP turns out to be a winner, particularly in the current scientific and technological scenario, which presents us with three important discontinuities with the past: a) a growing rate of technological innovation, b) an exponentially increasing availability of multimodal data, and c) a pressing demand for problem-oriented interdisciplinarity. This special issue is intended to illustrate the fertility of this scientific paradigm at the service of a few

challenging applications, covering the areas of Information Retrieval and Cultural Heritage.

The huge increase of digitized text data available on the web raises the complementary issue of making this information accessible to users in intelligent, language-aware ways. Although human inquiries are often highly selective, perspective-taking and personalized, careful and accurate classification of web pages is an essential precondition for building specialized repositories and focused crawlers. In their contribution to the present issue, Safae Lassri, El Habib Banlahmar and Abderrahim Tragha provide a comprehensive overview of the huge literature on the topic, based on up-to-date machine learning algorithms. They suggest that the increasing availability of as yet poorly exploited information, such as information deriving from the visual analysis of web pages, and more accurate filtering/removal of noisy and irrelevant content, are likely to pave the way to optimal classification accuracy in the near future.

A fundamental part of our Cultural Heritage has been passed down through generations in handwritten form. However, only a fragment of these original manuscripts is available in digital formats, preventing millions of people from first hand access and fruition. It is one of the most pressing challenges and responsibilities of our digital epoch to offer the possibility of describing, storing and accessing in digital formats hundreds of thousands of manuscripts that survived through time, with a view to their preservation, philological scrutiny and dissemination. A pool of researchers from the Italian National Research Council offers a broad, critical overview of some of the most challenging aspects of this process.

Pasquale Savino, Anna Tonazzini and Franca Debole report on a comprehensive data model and a cataloguing, storage and retrieval system for ancient documents. The framework presupposes the availability of high quality scans of ancient manuscripts and describes at some length the number of important and delicate steps to be taken not only for making this digitized material searchable and openly accessible, but also for augmenting its potential for research purposes, scholarly scrutiny or general information.

In fact, the digital approach to Cultural Heritage preservation is about creating new objects of scientific inquiry by multiplying information sources. Research specialists as well as the general public can nowadays benefit from a large array of technological tools providing several, interrelated overlays of information on top of the original documents. Anna Tonazzini, Pasquale Savino, Emanuele Salerno, Muhammad Hanif and Franca Debole discuss the use of digital technology for virtual restoration and visual analysis of degraded manuscripts. By focusing on the effects of “bleed-through” distortion, they illustrate the prospects of removing one noisy layer of spectral information while leaving other, more informative layers unaltered, to allow easier reading of the manuscript. Conversely, by exploiting non visible acquisition bands (such as Infrared and Ultraviolet) other layers of

contentful information, physically removed from their original support to make room for more recent writings (in so called *palimpsests*), may become readable once again.

Time introduces another multiplying dimension for ancient handwritten documents. More manuscripts of the same original text, made by different copyists, can contain so called variants, i.e. differences in the way the text was interpreted and transmitted by the copyists themselves. This turns a unitary textual material into a collection of different textual systems. The challenge for a philologist is, whenever possible, to restore the unity of the original text. Luigi Bambaci, Federico Boschetti and Riccardo Del Gratta describe a philological model for encoding the "critical apparatus" of a text, i.e. the collection of variant readings and corrections proposed by scholars. The challenge for applied digital technology is to provide an expressive, not too verbose encoding framework, i.e. a domain-specific meta-language, which can be interpreted and acted upon by a computer, for different views of the same text to be provided automatically, and for different information overlays to be classified and aligned with the original text.

Although only some of the contributions to the present issue directly discuss the use of machine learning technology in the context of a specific task, all of them emphasize the tremendous potential of this technology for providing assistance in the completion of repetitive tasks, or for anticipating and ranking possible solutions to the most common problems. The increasing availability of big repositories of multimodal data, advanced statistical techniques for data analysis and machine learning algorithms will enable scholars to investigate the interaction of a large number of factors affecting language usage in naturalistic contexts. Although we are still far away from having fully operational, virtual assistants in high-level intellectually challenging tasks like text interpretation, we expect NLP and machine learning algorithms to make human performance in the execution of these tasks considerably more reliable, consistent and error-free in a few years from now.

Acknowledgements
This special issue stems from the 3rd IEEE Conference on "Machine Learning and Natural Language Processing: Models, Systems, Data and Applications" held in Marrakech (October 21-24, 2018) within the framework of the Moroccan Chapter of IEEE CiSt'18. We would like to thank the Organizing Committee and the Program Committee of the Conference, and in particular Mohammed El Mohajir, Conference General Chair, for their constant support and encouragement. We are also grateful to the authors of the selected papers and all reviewers for their unflagging commitment and dedication. Thanks are also due to the co-editors in chief of IJIST for their patience, stamina and firm guidance.

Vito Pirrelli
Ouafae Nahli
Mohamed Chahhou

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D. & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language learning*, 59, 1-26

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4), 547-582.

Larsen-Freeman, D. and Cameron, L. (2008). *Complex Systems and Applied Linguistics*. Oxford, UK: Oxford University Press.

Libben, G. (2016). The quantum metaphor and the organization of words in the mind. *Journal of Culture Cognitive Science*, 1:49–55.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, 41(9), 1017-1023.