A Holistic Approach to Scientific Reasoning Based on Hybrid Knowledge Representations and Research Objects

Jose Manuel Gomez-Perez
Expert System
Madrid, Spain
jmgomez@expertsystem.com

Ronald Denaux
Expert System
Madrid, Spain
rdenaux@expertsystem.com

Andres Garcia
Expert System
Madrid, Spain
rdenaux@expertsystem.com

Raul Palma
PSNC
Poznan, Poland
rpalma@man.poznan.pl

1 MOTIVATION AND GOALS

Under the light of current developments in AI it appears the time is ripe for a shared partnership with machines, whereby humans can benefit from augmented reasoning and information management capabilities provided that machines are endowed with the necessary intelligence to assist with such tasks. This seems to be particularly the case of the scientific domain, where some envision the development of an AI that can make major scientific discoveries and that eventually becomes worthy of a Nobel Prize [9]. This vision may still be far from realization, but it is not completely new nevertheless.

NLP technologies based on well-formed, logically sound structured knowledge representations (knowledge graphs, ontologies) leverage expressive and actionable descriptions of the domain of interest through logical deduction and inference, and can provide logical explanations of reasoning outcomes. Closely related to this family of approaches, project Halo [7] aimed to develop a Digital Aristotle able to answer novel questions in scientific domains with expertise equivalent to Advanced Placement competence level. Halo enabled subject matter experts (SMEs) to model complex scientific knowledge from textbooks and related questions, based on an underlying logical formalism and a knowledge modeling workbench to assist SMEs in the task. The resulting system achieved an unprecedented question answering performance level for SME-entered knowledge, but it also had a number of severe drawbacks, including brittleness (coverage, precision or granularity gaps), scalability issues, and the need for a considerable force of well trained human labor to manually encode large amounts of scientific knowledge.

On the other hand, the last decade has witnessed a shift towards statistical methods due to the increasing availability of raw data and cheap computing power. These have proved to be powerful and convenient in many linguistic tasks, such as part-of-speech tagging or dependency parsing. However, they are also limited, e.g. humans seek causal explanations, which are hard to provide based on statistical induction rather than logical deduction. Recent results in the field of distributional semantics [10] have shown promising ways to learn features from text that can complement the knowledge already captured explicitly in structured representations. Embeddings provide a compact and portable representation of words and their meaning that stems directly from a document corpus. In this scenario, a notion of *semantic portability* [3] emerges that refers to the capability to capture as an information artifact (a vector) the semantics of a linguistic unit (a word) from its occurrences in the corpus and how such artifact enables that meaning to be merged with other forms of knowledge representation.

Furthermore, scientific knowledge is heterogeneous and can present itself in many forms. During its analysis phase, Halo produced an inventory of the different types of knowledge identified. Such knowledge types include among others: factual knowledge, procedural, classification, mathematical, diagrammatic, tabular and experimental. It is therefore clear that successfully reading and understanding scientific knowledge (either by humans or machines) requires addressing the different knowledge types in a holistic way, which remains a challenging task. We argue that addressing such challenge requires generalizing the notion of semantic portability from a text understanding scenario to a broader one where other modalities, such as diagrams, processes, experiments and related artifacts like scientific workflows and their execution provenance, are also involved. This can be achieved by learning individual models for each modality in the form of concept embeddings following a distributional semantics [8, 12] and learning the corresponding transformations between each vector space. The result will be a shared, hybrid formalism that encompasses the different modalities involved in scientific knowledge. Using embeddings to represent not only words but arbitrary features has been recently popularized by Chen and Manning in [2].

At this point, the question remains where to obtain the cross-modal data required to learn such models and the necessary transformations between them. We argue that the growing collections of research objects from different scientific disciplines available in repositories like ROHub.org [11] will play a key role in this regard. Conceptually speaking, a research object [1] is a container of scientific knowledge, a semantically rich aggregation of all the materials involved in a scientific investigation, such as papers and bibliography, numerical data, hypotheses, methods, experiments, workflows encoding such experiments and the provenance of their executions. A research object thus becomes the carrier of the scientific knowledge associated to a specific investigation. They also bring together all the necessary information to preserve scientific

K-CAP2017 Workshops and Tutorials Proceedings, 2017

work against potential decay [13] and can be shared, reused and cited in scholarly communications. As scholars move away from paper towards digital content, research objects have a key role to play in the way scientific results are communicated and validated by the communities, given the need for mechanisms that support the production of self-contained publications involving not only text but also data, methods and software implementations.

In [6], we show how research objects are key pieces of a human-machine scientific partnership. Building on that, we aim at furthering the role of research objects in such partnership, leveraging research object corpora of cross-modal scientific knowledge to develop hybrid models for scientific reasoning and question answering. During the workshop, we aim at sharing and discussing these ideas, explore related lines of work and establish areas of common interest and collaboration with the participants. Key topics and research questions we wish to address include: approaches for hybrid reasoning, question answering and explanation, methods to build portable knowledge representations of multimodal data, how to combine the knowledge extracted from each modality in the research objects to recompose a coherent, more complete view of the scientific facts documented by them, and how each modality interplay with each other in doing so.

2 ABOUT THE AUTHORS

This research is conducted by a team of researchers at Expert System's COGITO Lab and the Poznan Supercomputing and Networking Center. Through the years, we have developed a body of work in the intersection of several areas of AI that converge in the ideas discussed in this document, including NLP, Knowledge Discovery, Representation and Reasoning and new ways of scholarly communication and preservation of scientific knowledge (as research objects). This work aims at enabling machines to understand text and other modalities in which knowledge can be expressed in a way similar to how humans read, bridging the gap between both through semantically rich knowledge representations and human-machine interfaces. In doing so, we believe that such vision is best served through a combination of structured knowledge and probabilistic approaches. The main author of this document participated in project Halo as a member of the DarkMatter team, focused on process knowledge acquisition from textbooks and question answering by domain experts [4, 5]. He is also one of the founders and key personnel behind ROHub.org, the reference platform for research object management. ROHub currently hosts almost 2,500 research objects and 180 scientists in a variety of experimental and observational scientific disciplines like Biology, Astrophysics and Earth Science.

ACKNOWLEDGMENTS

This research is funded by the EU H2020 and national research projects EVER-EST (674907), xLiMe-ES (20160805) and DANTE (700367).

REFERENCES

[1] S Bechhofer, I Buchan, D De Roure, P Missier, J Ainsworth, J Bhagat, P Couch, D Cruickshank, M Delderfield, I Dunlop, M Gamble, D Michaelides, S Owen, D Newman, S Sufi, and C Goble. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29, 2 (2013), 599 – 611. https://doi.org/10.1016/j.future.2011.08.004 Special section: Recent advances in e-Science.

[2] Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 740–750. http://www.aclweb.org/anthology/D14-1082

[3] Ronald Denaux and Jose M Gomez-Perez. 2017. Towards a Vecsigrafo: Portable Semantics in Knowledge-based Text Analytics. In *Proceedings of the 2017 workshop on Hybrid Statistical Semantic Understanding and Emerging Semantics (HSSUES)*. CEUR Workshop Proceedings, Held in Conjunction with the 16th International Semantic Web Conference, Vienna, Austria.

[4] Jose Manuel Gomez-Perez, Michael Erdmann, Mark Greaves, and Oscar Corcho. 2013. A Formalism and Method for Representing and Reasoning with Process Models Authored by Subject Matter Experts. *IEEE Trans. on Knowl. and Data Eng.* 25, 9 (Sept. 2013), 1933–1945. https://doi.org/10.1109/TKDE.2012.127

[5] Jose Manuel Gomez-Perez, Michael Erdmann, Mark Greaves, Oscar Corcho, and V. Richard Benjamins. 2010. A framework and computer system for knowledge-level acquisition, representation, and reasoning with process knowledge. 68 (10 2010), 641–668.

[6] Jose M Gomez-Perez, Andres Garcia-Silva, and Raul Palma. 2017. Towards a Human-Machine Scientific Partnership Based on Semantically Rich Research Objects. In *eScience*. IEEE Computer Society, 1–9.

[7] David Gunning, Vinay K Chaudhri, Peter E Clark, Ken Barker, Shaw-Yi Chaw, Mark Greaves, Benjamin Grosof, Alice Leung, David D McDonald, Sunil Mishra, and Others. 2010. Project Halo Update—Progress Toward Digital Aristotle. *AI Magazine* 31, 3 (2010), 33–58.

[8] Zellig S. Harris. 1981. *Distributional Structure.* Springer Netherlands, Dordrecht, 3–22. https://doi.org/10.1007/978-94-009-8467-7_1

[9] Hiroaki Kitano. 2016. Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery. *AI Magazine* 37, 1 (2016), 39–49. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2642

[10] Tomác Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality.. In *NIPS*. https://doi.org/10.1162/jmlr.2003.3.4-5.951 arXiv:1310.4546

[11] Raul Palma, Piotr Hołubowicz, Oscar Corcho, Jose M Gomez-Perez, and Cezary Mazurek. 2014. ROHub—A Digital Library of Research Objects Supporting Scientists Towards Reproducible Science. In *Semantic Web Evaluation Challenge*. Springer, 77–82.

[12] Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics* 20, 1 (2008), 33–54.

[13] J Zhao, JM Gomez-Perez, K Belhajjame, G Klyne, E García-Cuesta, A Garrido, KM Hettne, M Roos, D De Roure, and C Goble. 2012. Why workflows break - Understanding and combating decay in Taverna workflows.. In *eScience*. IEEE Computer Society, 1–9. http://dblp.uni-trier.de/db/conf/eScience/eScience2012.html#ZhaoGBKGGHRRG12