Towards a Practical Emergent Knowledge Exploitation

Towards a Practical Emergent Knowledge Exploitation

Vít Nováček
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
E-mail: FirstName.LastName@deri.org

1 Problem Description

My PhD is about researching novel means for representation and exploitation of emergent knowledge extracted from free text. The aim is to provide a more applicable alternative to state-of-the-art KR approaches that are rather insufficient in this context. The primary motivating use case is knowledge-based intelligent search in life science literature. Both lack and necessity of respective solutions have recently been confirmed by our discussions with medical practitioners. Their opinion seems to be shared by a broader community of publishers and researchers involved in the Elsevier Grand Challenge competition (cf. http://www.elseviergrandchallenge.com/). Thus, the thesis topic is much rather driven by the *market pull* than attempting for a *technology push*.

The problem addressed by the thesis has several aspects, as analysed in the remainder of this section. Large scale manual acquisition of knowledge that could be exposed to users of prospective "intelligent" systems is hardly possible. Therefore, (semi)automated solutions are considered as a way to go [Bechhofer and others, 2003]. One of the most prominent automated approaches is ontology learning from text [Maedche and Staab, 2004]. The automatically extracted emergent knowledge is dynamic, often explicitly vague, potentially inconsistent and incorrect, though [Haase and Völker, 2005]. This hampers its meaningful exploitation by the traditional, mostly brittle logics-based querying and inference frameworks [Bechhofer and others, 2003].

Approaches handling noisiness and uncertainty within traditional KR&R generalisations have been proposed – e.g., in [Straccia, 2006]. However, a practical, robust and general-purpose modelling covering the emergent knowledge is deemed hardly possible using the prevalent formal methods [Sheth *et al.*, 2005]. Even if the noisiness of learned ontologies was remedied before applying traditional methods for their exploitation, the shallow structure of the ontologies does not allow for many non-trivial conclusions, anyway [Haase and Völker, 2005; Bechhofer and others, 2003].

Examining alternative inspirations, recent advances in neuroscience [Frith, 2007] show that the most robust inference engines we know of—people—do not actually use logical principles in order to cope with noisy and incomplete information efficiently. [Frith, 2007] and [Gentner *et al.*, 2001] suggest that empirical similarity-based inference is more appropriate in practical settings involving noisy and uncertain data. The thesis essentially rests on applying these inspirations to a novel way of emergent knowledge representation and exploitation.

2 Solution Outline and Research Plan

Regarding syntax, the proposed knowledge representation de facto builds on a simple, standard and widely used triple model (cf. http://www.w3.org/TR/rdf-primer/). However, the subject-predicate-object statements are extended by heuristic positive or negative certainty measures and organised in so called conceptual matrices concisely representing every positive and negative relation of an entity to other entities. Soft empirical updates of the content are realised by fuzzy matrix aggregation operators. Metrics can be easily defined on the conceptual matrices, serving as a natural basis for gradual concept similarities [Nováček, 2009]. These are employed by light-weight inference services of two basic types: 1. *retrieval* of similar concepts (quite straightforward); 2. fixed-point *rule-based materialisation* of implicit relations, complex *querying* (similarity as a basis for soft variable unification and for approximate fixed-point computation). The inference algorithms have anytime behaviour and it is possible to programmatically adjust their completeness/efficiency trade-off.

The suggested solution supports precise legacy knowledge bases that can serve as seed models refining the automatically extracted noisy facts. Editing and querying interfaces with minimal learning requirements (i.e., close to natural language) are supposed to encourage involvement of domain experts in further emergent knowledge refinement. To facilitate user satisfaction, the results of queries on the emergent knowledge bases have to be sorted according to both absolute and query-specific relevance.

The tentative thesis title is *Towards a Practical Emergent Knowledge Exploitation*. It will contain—in line with the adopted research plan—the following material: (i) motivation, inspirations and comparison to the state-of-the-art; (ii) technical description of the solution; (iii) formal analysis of the implemented algorithms, their anytime behaviour and customisable completeness/efficiency trade-off; (iv) architecture of the respective research prototype and detailed description of its implementation; (v) extensive technical evaluation of the prototype's precision/recall w.r.t. to a gold stan-

dard created by domain experts, comparison to base-line approaches (RDFS inference, an inconsistency-tolerant and an uncertain logical reasoner), performance/scalability analysis, user-based applicability assessment; (vi) outline of the industrial potential of the delivered results based on the evaluation. Particular self-contained results are to be published as appropriate conference or journal contributions.

3 Progress to Date and Remaining Tasks

By mid 2008, I have firmly settled on my topic and started to elaborate the basic principles of the proposed emergent knowledge representation framework and inference services. In the end of 2008, I finished a proof-of-concept implementation of the proposed knowledge representation and basic inference services, called EUREEKA (a permutated acronym for *Easy to Use Empirical Reasoning about Automatically Extracted Knowledge*). Details on the framework, recent updates, promising preliminary results and related approaches are described in [Nováček, 2009; Nováček *et al.*, 2009].

To outline the typical EUREEKA work-flow, the emergent knowledge processing makes use of a so called *ACE* pipeline (the pipeline can be executed even as ACE*, i.e., as a search for a global fixed point of the respective operations):

- *A* for *addition* – the extracted knowledge is incrementally added into a seed knowledge base (either empty, or initialised with a legacy ontology serving for soft emergent knowledge refinement)
- *C* for *closure* – after the addition batch, we compute the closure of the actual content using an approximate fixed-point evaluation of a domain rule-set
- *E* for *extension* – the extracted concepts are analogically extended by means of similar, yet richer targets in the knowledge base

The resulting knowledge base is exposed to users using a simple conjunctive query language with negation. The answer statements are sorted according to their relevance within the knowledge base – see [Nováček *et al.*, 2009; Nováček, 2009] for details. Examples of queries and selected answers are as follows:

`QUERY: ? : type : breast cancer` $\rightsquigarrow$ `ANSWER: <cysto-sarcoma phylloides : TYPE : breast cancer>`^{1} `...QUERY: rapid antigen testing:part of:? AND ?:type:clinical study` $\rightsquigarrow$ `ANSWER: <dicom study : USE : protein info>`$^{0.8}$ `AND <initial study : INVOLVED : patients>`$^{0.912}$ `...QUERY: acute granulocytic leukemia: NOT type :T-cell leukemia` $\rightsquigarrow$ `ANSWER: <acute granulocytic leukemia : TYPE : T-cell leukemia>`$^{-0.664}$ `...`

As of March 2009, EUREEKA has already been deployed as a crucial building block of CORAAL [Nováček *et al.*, 2009], a knowledge-based life science literature search engine. CORAAL exposes (in real time) more than 15 millions of automatically extracted and inferred fuzzy statements coming from about $12,000$ cancer research publications. The system has been internationally recognised as a recent finalist of the Elsevier Grand Challenge competition.

Several important tasks still remain to be accomplished, though. 1. Formal properties of the algorithms have to be properly analysed – especially the influence the parameters adjusting the convergence speed have on the completeness of the eventual result. 2. The "philosophical" and theoretical relationships between the implemented solution and current approaches to knowledge representation have to be investigated deeper for the thesis. 3. Both technical and user-based preliminary evaluation described in [Nováček, 2009] and [Nováček *et al.*, 2009], respectively, have to be rigorously extended in line with the afore-mentioned research plan. Further work on dissemination of the recent and future results at appropriate venues is also needed in order to prove the thesis potential.

Acknowledgments The PhD work has been supported by the EU IST 6th framework's projects 'Knowledge Web', 'Nepomuk' (FP6-507482, FP6-027705) and the 'Líon', 'Líon II' projects funded by Science Foundation Ireland (SFI/02/CE1/I131, SFI/08/CE/I1380).

References

[Bechhofer and others, 2003] Sean Bechhofer et al. Tackling the ontology acquisition bottleneck: An experiment in ontology re-engineering, 2003. At `http://tinyurl.com/96w7ms`, Apr'08.

[Frith, 2007] Chris Frith. *Making Up the Mind: How the Brain Creates Our Mental World*. Blackwell, 2007.

[Gentner *et al.*, 2001] Dedre Gentner, Keith J. Holyoak, and Boicho K. Kokinov, editors. *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press, 2001.

[Haase and Völker, 2005] Peter Haase and Johanna Völker. Ontology learning and reasoning - dealing with uncertainty and inconsistency. In *Proceedings of the URSW2005 Workshop*, pages 45–55, NOV 2005.

[Maedche and Staab, 2004] Alexander Maedche and Steffen Staab. Ontology learning. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, chapter 9, pages 173–190. Springer, 2004.

[Nováček *et al.*, 2009] Vít Nováček, Tudor Groza, and Siegfried Handschuh. Towards knowledge-based life science publication repositories. In *Semantic e-Science*. Springer Verlag, 2009. In press. Pre-print available at `http://tinyurl.com/d2gjv5`.

[Nováček, 2009] Vít Nováček. Towards an efficient knowledge-based publication data exploitation: An oncological literature search scenario. Technical Report DERI-TR-2009-03-23, DERI, NUIG, 2009. Available at `http://tinyurl.com/csh3rf`.

[Sheth *et al.*, 2005] Amit Sheth, Cartic Ramakrishnan, and Christopher Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *Int'l Journal on Semantic Web & Information Systems*, 1(1):1–18, 2005.

[Straccia, 2006] Umberto Straccia. A fuzzy description logic for the semantic web. In Elie Sanchez, editor, *Fuzzy Logic and the Semantic Web*, Capturing Intelligence, chapter 4, pages 73–90. Elsevier, 2006.