Why the Semantic Web Has Never Got Too Much of a Meaning and How to Put It There*

Vít Nováček, Siegfried Handschuh, Stefan Decker

Digital Enterprise Research Institute (DERI), National University of Ireland Galway
E-mail: vit.novacek@deri.org

Abstract. Our outrageous idea concerns the very core of the Semantic Web – the (lack of) semantics themselves. In Section 1 we claim that the presently used conception of web semantics fails to capture the link between the machine-readable descriptions and their actual meaning (in the sense of a formal grounding in reality). We also argue that this is inherently related to culprits of various difficulties the community has been coping with (e.g., knowledge acquisition bottleneck or data integration challenges). In Section 2 we propose a remedy of this unsatisfactory state of affairs – a broadened notion of emergent web meaning that can be derived in a bottom-up manner from the web data using principles of distributional semantics. Section 3 gives an overview of a preliminary solution [1] we recently implemented with promising results. Finally, we outline directions of future research required to fully realise our vision.

1 The (Lack of) Semantics in the Semantic Web

The Semantic Web has been designed for asserting meaning to things via machine-readable specifications – knowledge bases composed of formal symbolic statements (e.g., RDF descriptions or OWL ontologies). Due to numerous successful research and development efforts in the last decade, we know a great deal about how to query, classify, materialise, analyse or interpret such knowledge bases. However, mechanisms for their formal grounding in reality are largely unexplored, which makes the meaning of the Semantic Web effectively incomplete.

The main problem is that current approaches consider the basic symbols (URIs) occurring in the Semantic Web expressions as atomic and monolithic units of meaning. If this was true, though, we can interpret URIs absolutely arbitrarily and the Semantic Web will still be 'meaningful'. This is obviously ridiculous, as the URIs actually do have a deeper and non-arbitrary meaning. However, this meaning is induced by a distributed and consensual agreement of agents on the web. As such, it is inherently fluid and unstable (it changes in time and across application domains and/or user communities). This is rather difficult to capture by the current symbolic approaches to the web semantics, therefore the meaning of URIs is conceived as a sort of 'black-box' we do not need to care

* This work has been supported by the 'Líon II' project funded by SFI under Grant No. SFI/08/CE/I1380.

about. Such an approach results in omitting the important link between the Semantic Web expressions and their grounding in the reality (i.e., their meaning in the traditional sense of works like [2]). A bottom-up and empirically verifiable assignment of meaning to the symbolic descriptions on the Semantic Web would allow for much more complete and plausible semantics. Unfortunately, we currently do not know how to do that in a well-founded way, which leads to a number of issues related both to foundations and practical applications of the Semantic Web.

The fact that the semantics of Semantic Web are hardly actual semantics may trouble few perfectionist theorists, but not so much the application developers and companies. Yet we believe the incompleteness of the Semantic Web's meaning is also a fundamental culprit of many seemingly unrelated problems that trouble us nowadays due to their tangible economic impact. One example is the *knowledge acquisition bottleneck*. Creating expressive machine-readable descriptions of the web reality is expensive, while their much cheaper automated generation is usually too inaccurate [3]. The main reason is that the relation between the real world phenomena and their formal descriptions is very difficult to be reliably captured with the current approaches. Thus it requires either substantial collaborative efforts, or often undesirable level of simplification. The acquisition problem is closely related to *imprecision* – without a well-founded, non-arbitrary link between the reality and its symbolic descriptions, it is difficult to achieve and/or validate the precision and actuality of the Semantic Web knowledge bases, which hampers their reliability and applicability. Then there is *contextuality*. Even the most immaculately crafted knowledge representations depend on several types of context – the domain, motivations of the creators, the current state of the evolving background knowledge, etc. A comprehensive representation of the contextual dependence is rather problematic in the current Semantic Web, though, again arguably due to the unclear grounding of the descriptions in the reality (and context). *Data integration* is also hampered by the arbitrary nature of the web semantics. Ontologies that are supposed to be used for the data integration are available for a lot of domains, however, one often has to perform a non-trivial ontology alignment before proceeding with the data integration itself [4]. Integration based on the empirical meaning induced by the data itself could elegantly circumvent these difficulties.

2 Towards a Remedy

We conjecture that the solution of all the afore-mentioned problems consists simply of changing our views on what the semantics of the web actually are. If providing expressive, accurate, empirically grounded machine-readable descriptions of the web is apparently so hard and/or impractical, how about giving up on providing anything complex straightaway? How about focusing on something easier first, like naming the things on the web and describing simple relations between them? This sounds familiar. It is exactly what has been happening within the Linked Open Data initiative that has recently gained so much momentum.

Still, this is only a first step. Naming things and links between them by itself will not help us to capture many interesting semantic phenomena. It will only provide us with elementary particles from which actual patterns of web-based knowledge are being composed in a fluid, continuous and emergent process. We have to stop thinking about the Semantic Web as an effort to *assert* the machine-readable annotations of the web data in a top-down manner. We have to start finding new techniques for *observing* bottom-up patterns that emerge from the complex system of the interactions on the web of data. Such patterns can have many forms – they can be clusters of implicitly related entities, generalised relationships between classes of similar entities, delimitations of subsets of web data defining particular domain contexts, or formalised analogies. From the observed patterns, more concise and expressive symbolic descriptions can be constructed and processed further, however, now they will not be arbitrary anymore, as they have provably originated from the reality (or its most comprehensive approximation currently accessible to machines – the web).

The focal points of our new approach to the web semantics are: (1) *production* of simple elementary statements from the extant data (e.g., the contents of the 'human' web, databases or linked data cloud); (2) *extraction* of complex and general patterns from the elementary statements; (3) *interpretation* of the extracted patterns; (4) *application* of the extracted and interpreted patterns. Note that all these steps can easily be made compatible with the basic and commonly used Semantic Web standards, namely RDF. The elementary statements can very well be represented as RDF triples, while the complex patterns generated from large bases of simple triples can be rendered as RDF statements again during the interpretation step. The results of this processes can then be added to the original datasets to be further utilised (materialised, queried, visualised, etc.) by the existing technologies.

When all the steps of the suggested approach are automated with a sufficient reliability and coverage, we get a powerful new arsenal to cope with the problems mentioned in Section 1. For instance, simple relationships are much easier to be automatically learned from legacy data (like text or databases) than complex logical axioms. This and the consequent extraction of more interesting general patterns from the simple statements can alleviate the knowledge acquisition bottleneck. Similarly, imagine that the machine-readable knowledge about a domain can be inexpensively generated for any snapshot of domain data anytime. Then one can substantially reduce the need for technologically and economically demanding maintenance that is necessary to keep the current Semantic Web representations precise and up to date.

3 Preliminary Solution and Further Steps

In our ISWC'11 research track paper [1], we introduced a basic implementation of the notion of semantics we advocate here. The approach in [1] stems from recent advances in distributional semantics. This sub-field of computational linguistics is based on a hypothesis that "a word is characterized by the company it

keeps" [5]. In our context, we can rephrase this to characterise the meaning of a thing by the company of things linked to it. In order for such meaning to be representative, though, we have to analyse the 'company' across as much content as possible. To do so, we employed an approach utilising simple, yet universal and powerful tensor-based representation of distributional semantics proposed in [6], which we adapted to the Semantic Web specifics. We also showed how one can execute rules on the top of the tensor representation, which effectively leads to a coherent combination of the bottom-up (distributional) and top-down (symbolic) approaches to meaning. The resulting framework can be used for clustering of related entities or properties, semantic search, incremental induction and execution of rules, or discovery of analogies among Semantic Web data. These tasks are rather varied, yet with our approach, one can tackle them coherently by applying well-founded linear algebra and statistical data analysis methods to different perspectives of the underlying tensor representation. Details, overview of related work and evaluation of the approach are provided in [1].

Although we may have already provided a first proof of concept and a promising practical application of the introduced approach to web semantics, a lot of work remains to be done in order to fully realise the potential of our proposal. First of all, we need to refine formalisms for representation and exploitation of the distributional, emergent web semantics. Another field of research are improved techniques for extraction of the elementary statements from legacy data, and for inference of a wider range of complex distributional patterns from the elementary statements. The combination of the symbolic (top-down) and distributional (bottom-up) approaches to semantics requires a deeper investigation, too. In order to ensure truly web-scale applicability of the introduced principles, parallelised and distributed implementations of the required data storage and processing techniques have to be devised. Last but not least, appropriate methods for interactive presentation of the emergent web knowledge to users must be researched. This is also related to interweaving the social and Semantic Web, which will facilitate direct participation of people in the evolution of the emergent web knowledge. This plan will not be easy to realise, but we believe that walking these thorny paths can be an exciting endeavour that will eventually lead to much better exploitation of the still largely dormant web potential.

References

1. Nováček, V., Handschuh, S., Decker, S.: Getting the meaning right: A complementary distributional layer for the web semantics. In: Proceedings of ISWC'11, Springer (2011) In press. Pre-print available at: `http://goo.gl/FRT77`.
2. de Saussure, F.: Course in General Linguistics. Open Court, Illinois (1983)
3. Bechhofer, S., et al.: Tackling the ontology acquisition bottleneck: An experiment in ontology re-engineering (2003) At `http://tinyurl.com/96w7ms`, Apr'08.
4. Euzenat, J., Shvaiko, P.: Ontology matching. Springer (2007)
5. Firth, J.: A synopsis of linguistic theory 1930-1955. Studies in Ling. Anal. (1957)
6. Baroni, M., Lenci, A.: Distributional memory: A general framework for corpus-based semantics. Computational Linguistics (2010)