SHELDON: Semantic Holistic framEwork for LinkeD ONtology data

Diego Reforgiato Recupero[1], Andrea Giovanni Nuzzolese[1], Sergio Consoli[1], Aldo Gangemi[1,2], and Valentina Presutti[1]

[1] STLab-ISTC Consiglio Nazionale delle Ricerche, Catania, Italy
[2] LIPN Sorbonne-Cité, Université de Paris-13.

Abstract. SHELDON is the first true hybridization of NLP machine reading and Semantic Web. It is a framework that builds upon a machine reader for extracting RDF graphs from text so that the output is compliant to Semantic Web and Linked Data patterns. It extends the current human-readable web by using Semantic Web practices and technologies in a machine-processable form. Given a sentence in any language, it provides different semantic functionalities (frame detection, topic extraction, named entity recognition, resolution and coreference, terminology extraction, sense tagging and disambiguation, taxonomy induction, semantic role labeling, type induction, sentiment analysis, citation inference, relation and event extraction) as well as nice visualization tools which make use of the JavaScript infoVis Toolkit and RelFinder, as well as a knowledge enrichment component that extends machine reading to Semantic Web data. The system can be freely used at `http://wit.istc.cnr.it/stlab-tools/sheldon`.

1 Introduction

The Machine Reading paradigm relies on bootstrapped, self-supervised Natural Language Processing (NLP) performed on basic tasks, in order to extract knowledge from text. Machine reading is typically much less accurate than human reading, but can process massive amounts of text in reasonable time, can detect regularities hardly noticeable by humans, and its results can be reused by machines for applied tasks. SHELDON performs a hybrid (part of the components are trained, part are rule-based), self-supervised variety of machine reading that generates RDF graph representations out of the knowledge extracted from text by tools dedicated to basic NLP tasks. Such graphs extend and improve NLP output, and are typically customized for application tasks.

SHELDON includes and extends several software components successfully evaluated in the recent past [15,6,5,7,16,9,14,2,8,10].

The machine reading capability of SHELDON is based on FRED [15,5], a tool for automatically producing RDF/OWL ontologies and linked data from text. FRED integrates, transforms, improves, and abstracts the output of several NLP tools. The backbone deep semantic parsing is currently provided by

Boxer [3] which uses a statistical parser (C&C) producing Combinatory Categorial Grammar trees. Several heuristics are adopted in order to exploit existing lexical resources and gazetteers to generate representation structures according to Discourse Representation Theory (DRT), which generates formal semantic representation of text through an event (neo-Davidsonian) semantics. The basic NLP tasks performed by Boxer, and reused by FRED, include: event detection (FRED uses DOLCE+DnS[3] [4]), semantic role labeling with VerbNet[4] and FrameNet roles, first- order logic representation of predicate-argument structures, logical operator scoping (called boxing), modality detection, tense representation, entity recognition using TAGME[5], word sense disambiguation (the next version is going to use BabelNet[6]), DBpedia for expanding tacit knowledge extracted from text, etc. All is integrated and semantically enriched in order to provide a Semantic Web-oriented reading of a text.

Revealing the semantics of hyperlinks has a high potential impact on the amount of Web knowledge that can be published in machine readable form, keeping the binding with its corresponding natural language source. SHELDON integrates LEGALO [14], a novel method for uncovering the intended semantics of links by tagging them with semantic relations. LEGALO implements a set of graph pattern-based rules for extracting, from FRED graphs, Semantic Web binary relations that capture the semantics of specific links.

SHELDON is able to give a boost to the sentiment analysis practice. One of its components is built on top of SENTILO [16,7], a domain-independent system that performs sentiment analysis by hybridizing natural language processing techniques and Semantic Web technologies. Given a sentence expressing an opinion, SENTILO recognizes its holder, detects the topics and subtopics that it targets, links them to relevant situations and events referred to by it and evaluates the sentiment expressed on each topic/subtopic. It uses an ontology for opinion sentences, a new lexical resource that enables the evaluation of opinions expressed by means of events and situations, and a novel scoring algorithm for opinion sentences.

SHELDON also performs definitional taxonomy induction, integrating its result into the RDF graph of the text. The component for this task is based on TÌPALO [6]. TÌPALO identifies the most appropriate types for an entity by interpreting its natural language definition, which is extracted from its corresponding Wikipedia page abstract. TÌPALO relies on FRED for parsing and representing the logical form of a given sentence and induce a taxonomy by reusing WordNet types, WordNet supersenses, and DUL types.

Bibliographic citations are the most used tools of academic communities for linking research, for instance by connecting scientific papers to related works or source of experimental data. SHELDON implements a strategy for the linking within scientific research articles feature based on CITALO [9], a tool to infer

[3] D. U. L. Ontology. http://www.ontologydesignpatterns.org/ont/dul/dul.owl

[4] T. V. project. http://verbs.colorado.edu/mpalmer/projects/verbnet.html

[5] http://tagme.di.unipi.it/

[6] http://babelnet.org/

automatically the function of citations by means of Semantic Web technologies and NLP techniques. A sentence containing a reference to a bibliographic entity and the CiTO [12] ontology used to describe the nature of citations in scientific research articles are taken as input to infer the function of that citation. CITALO relies on FRED to extract ontological information from the input sentence.

Besides the graph visualization (displayed using Graphviz[7]), and the triple output for each component of SHELDON, there is also another data visualization component which is built on top of the Semantic Scout [2], which uses the JavaScript InfoVis Toolkit[8]. Finally, it is possible to augment the identified relations between detected DBpedia entities using a SHELDON component that is built starting from the expansion algorithm described for RelFinder [8] and that shows those relations using the RelFinder graphical interface.

SHELDON provides a REST API for each of its components so that everyone can build online end-user applications that integrate, visualize, analyze, combine and infer the available knowledge at the desired level of granularity. Potentially, each stakeholder interested in semantic aggregate information for multilingual text could be a customer. A start up is going to be founded in UK which will exploit SHELDON's technology (with only commercially-viable components) as one of its main cutting-edge products (we are currently solving some licensing issues).

2 SHELDON at work

Fig. 1(a) shows the main interface of SHELDON where one can type a sentence in any language and decide which semantic task to perform. The reader will notice that SHELDON will always provide the results in English. The Bing Translation APIs[9] have been used and embedded within SHELDON.

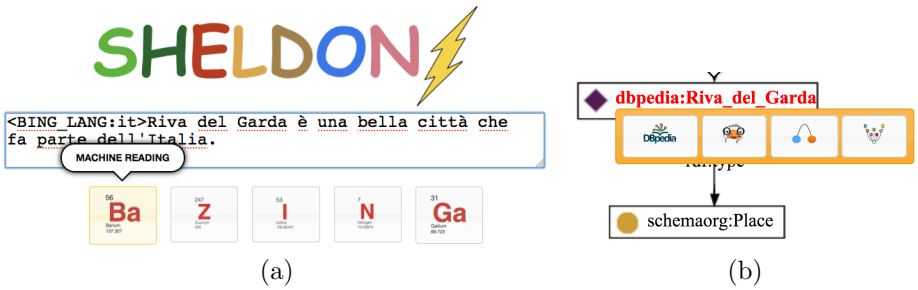

Fig. 1: SHELDON front page (a), SHELDON's navigation toolbar for identified DBpedia entities(b)

If the used language of the sentence is different than English, then the tag *<BING_LANG:lang>* needs to precede the sentence, where *lang* indicates a code

[7] Graphviz - Graph Visualization Software, http://www.graphviz.org/

[8] http://philogb.github.io/jit/

[9] http://www.microsoft.com/web/post/using-the-free-bing-translation-apis

for the language of the current sentence[10]. For example, the sentence:

<BING_LANG:it>Riva del Garda è una bella città che fa parte dell'Italia[11].

would be a valid Italian sentence to be processed.

SHELDON addresses the following four main tasks:

- If the machine reading capability is chosen (by pressing the button "Ba" of the toolbar - cf. Figure 1(a)), SHELDON would output an RDF graph with several associated information (detected DBpedia entities, events and situations mapped within DOLCE, WordNet and VerbNet mapping, pronoun resolution).
- If the sentiment analysis option is used (by pressing the button "Z" of the toolbar), for the same sentence SHELDON would return a RDF graph annotated with concepts from a sentiment analysis ontology with scores for the positive adjective, *beautiful*, and with an opinion score for the entity *Riva del Garda* which was computed according to the sentiment propagation algorithm discussed in [16].
- If the relation discovery choice is made (by pressing the button "I" of the toolbar), SHELDON returns the new (or aligned against existing repositories) relations between DBpedia entities in the text, in our example *Riva del Garda* and *Italy*.
- If the citation typing is chosen (by pressing the button "N" of the toolbar), SHELDON processes the sentence as a citation context, i.e., a piece of text containing an explicit citation (marked as "[X]" in the text) to a scholarly article. In this case SHELDON infers automatically the function of a citation by means of Semantic Web technologies, NLP and Sentiment Analysis techniques. The output is a property of the CiTO ontology[12], which provides a set of 41 properties for describing the nature of citations in scientific research articles and other scholarly works.

Except for the citation typing task (which returns a single property of the CiTO ontology), the SHELDON's user interface returns an interactive RDF graph that can be used by an user for browsing the knowledge resulting from the specific task. When clicking on each DBpedia entity node displayed in a graph, a pop-up menu appears (cf. Figure 1(b)). This menu allows an user to perform different actions, namely:

- visualization of an entity's page on DBpedia;
- exploratory search with Aemoo[13] [10] starting from a given entity;
- relation augmenting with [14], which allows to discover new relations for a DBpedia entity by exploiting the outgoing links and the natural language available from the corresponding Wikipedia article;

[10] check `http://msdn.microsoft.com/en-us/library/hh456380.aspx` for the list of language codes.

[11] The English translation is "Riva del Garda is a beautiful city which is part of Italy".

[12] CiTO: `http://purl.org/spar/cito`. [13]

[13] Aemoo: `http://www.aemoo.org`

– typing information augmenting with TÌPALO [6], which returns the most appropriate types for a given DBpedia entity by analyzing the natural language available from the corresponding Wikipedia abstract for such an entity. In particular, by clicking on the DBpedia entity, a new RDF graph composed of `rdf:type`, `rdf:subclassOf`, `owl:sameAs`, and `owl:equivalentTo` statements providing typing information is returned. These are aligned to the DBpedia Ontology, WordNet 3.0 in RDF, DUL, and DolceZero.

Fig. 2(a) shows the produced output of the machine reader feature, Fig. 2(b) shows the output for the semantic sentiment analysis whereas Fig. 2(c) shows the produced relation between the DBpedia entities recognized within the sentence.

For the machine reading, the sentiment analysis and the relation finding capabilities it is even possible to show the complete list of RDF triples (syntactic constructs, offset between words and input sentence, URIs of recognized entities, text span markup specification support using Earmark [11], relations between source and translated text) that SHELDON outputs by choosing a view (RDF/XML, RDF/JSON, Turtle, N3, NT, DAG) other than the Graphical View item which is set by default.

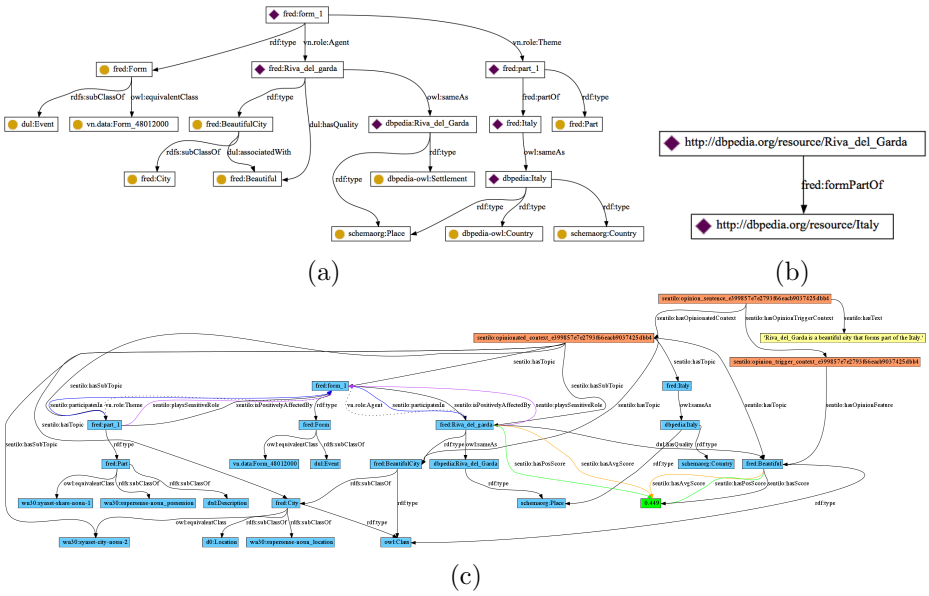

Fig. 2: Machine reader output (a), semantic link identification (b), and semantic sentiment analysis output (c) for the Italian sentence *Riva del Garda è una bella città che fa parte dell'Italia*

Within the options at the bottom of the produced graphs it is possible to export the graph as a PNG or JPEG image, to see the augmented knowledge for the identified DBpedia entities from SHELDON using a nice GUI built on top of RelFinder and to navigate the graph through a nice visualization tool that builds upon the Semantic Scout [2] and that uses the JavaScript InfoVis Toolki.

A Addressing Semantic Web Challenge Requirements

Table 1 describes how SHELDON addresses minimal requirement specified by the Semantic Web Challenge call. SHELDON addresses also a number of desirable features which are described in Table 2.

The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.
SHELDON has a practical value to both general Web users and application developers. General Web users can appreciate the graph representation of a given sentence using the visualization tools provided, semantics expert can see the RDF triples in more detail, researchers can test the automatic annotations provided for citations within research papers and build a semantic map of relevant articles within a certain domain, and application developers can use the REST API for empowering applications using its different capabilities. Developers of intelligent applications (e.g. semantic web expert systems) can use SHELDON for several tasks and understand the results better though the graph representation visualization.
The information sources used should be under diverse ownership or control, should be heterogeneous (syntactically, structurally, and semantically), and should contain substantial quantities of real world data (i.e. not toy examples).
We use an extension of VerbNet in RDF (currently under our control), and WordNet-RDF [17] as lexical linked data, Stanford CoreNLP[a], DBpedia as world linked data, DUL, WATSON[b], LOV[c], NELL[d], TAGME, RelFinder[e] SentiWordNet [1] and SenticNet[f]. SHELDON can be used for processing any natural language resource on the web in any language and can link the extracted entities and concepts to the above mentioned resources according to linked open data best practices.
The meaning of data has to play a central role. Meaning must be represented using Semantic Web technologies. Data must be manipulated/processed in interesting ways to derive useful information and this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;
SHELDON semantics catches the meaning of diverse resources: factual data from either structured or unstructured sources, lexical data, and sentiment data, by reusing RDF as a common representation language and OWL for the ontologies that homogeneously describe the data and the opinion model (triggers, holders, topics, etc.). Data are manipulated in non-trivial ways: factual and lexical knowledge are tightly coupled based on a semiotic meta-model (cf. FRED paper [15]), SPARQL querying is extensively used to morph FRED graphs according to the related ontology model considered and using novel visualization techniques for easy visualization of RDF triples. Without meta-modeling and Semantic Web standards it would be very hard to reproduce SHELDON results.

[a] `http://nlp.stanford.edu/software/corenlp.shtml`

[b] `http://watson.kmi.open.ac.uk/WatsonWUI/`

[c] `http://lov.okfn.org/dataset/lov/`

[d] `http://rtw.ml.cmu.edu/rtw/`

[e] `http://www.visualdataweb.org/relfinder.php`

[f] `http://sentic.net/`

Table 1: Addressing Minimal Requirements

References

1. A. E. S. Baccianella and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, may 2010.

The application provides an attractive and functional Web interface (for human users)
SHELDON provides a very intuitive and google-like interface where it is possible to choose the task (machine reader, semantic sentiment analysis, relation discovery, definition of citations, relationship augmentation between identified DBpedia entities) to perform. The advanced navigation tools of SHELDON makes use of RelFinder and the JavaScript InfoVis Toolkit to create interactive data visualizations for the web.
The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use all data that is currently published on the Semantic Web.
SHELDON capabilities use quite complex pipeline, taking data from NLP machine reading components, interpreting them as RDF graphs, linking them to existing linked data, identifying situations/events, reinterpreting them as opinionating/opinionated content, identifying relations among research paper, augmenting results with further knowledge coming from external ontologies.
Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.
Each tool that SHELDON uses has been successfully demonstrated with results published in top international journals and conferences [15,6,5,7,16,9,14,2].
Novelty, in applying semantic technology to a domain or task that have not been considered before
Each component of SHELDON is novel and applied to new domains. The automatic extraction (or alignment with existing ones) of relations between entities in a given sentence for example is a new feature in the Semantic Web area. The exploitation of Semantics within the Sentiment Analysis area is also a new task which started with the Sentic Computing[a] where Sentiment Analysis is getting a boost from semantic technology. The automatic identification of the nature of citations of scholarly papers is performed by applying a novel approach, which relies on machine reading and sentiment analysis and, to the best of our knowledge has not been investigated yet in the area of semantic publishing.
Functionality is different from or goes beyond pure information retrieval
Each component of SHELDON applies machine reading, Semantic Web best practices based on the adoption of Ontology Design Patterns and Linked Data principles, cognitive computation, relation extraction, frame detection, automatic typing of entities and the automatic labeling of citation functions.
Multimedia documents are used in some way
SHELDON performs machine reading from natural language text, which can be provided in a variety of different multimedia formats, i.e., plain text, HTML, PDF and Word documents (we are currently working on the last three formats). Additionally, SHELDON supports speech recognition by relying on the Web Speech API[b]. Currently, the speech recognition in enabled by requesting an user to pronounce a sentence via the Web interface, but, for the future work, we are planning to extend SHELDON in order to accept audio files.
There is support for multiple languages and accessibility on a range of devices
SHELDON leverages BING[c] service and allows input sentences given in any language covered by BING (there are 44 different languages). The translation service has been embedded in the pipeline optimizing the remote calls and distributing them in parallel for quicker access.
The application has clear commercial potential and/or large existing user base
SHELDON is a killer application that can persuade a critical mass of business users to exploit its components for a wide variety of tasks (sentiment analysis, automatic typing of entities, knowledge extraction, automatic identification of relations between entities in a given text, text summarization, system recommendation, research papers citations annotation.

[a] `http://sentic.net/sentics/`

[b] Speech recognition is supported only with the Chrome browser. The documentation about the Web Speech API is available at `https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html`.

[c] `http://www.bing.com/translator/`

Table 2: Addressing Additional Desirable Feature

2. C. Baldassarre, E. Daga, A. Gangemi, A. M. Gliozzo, A. Salvati, and G. Troiani. Semantic scout: Making sense of organizational knowledge. In P. Cimiano and H. S. Pinto, editors, *EKAW*, volume 6317 of *Lecture Notes in Computer Science*, pages 272–286. Springer, 2010.
3. J. Bos. Wide-coverage semantic analysis with boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 277–286, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
4. A. Gangemi. What's in a schema? *Cambridge University Press, Cambridge, UK*, pages 144–182, 2010.
5. A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *ESWC*, volume 7882 of *Lecture Notes in Computer Science*, pages 351–366. Springer, 2013.
6. A. Gangemi, A. G. Nuzzolese, V. Presutti, F. Draicchio, A. Musetti, and P. Ciancarini. Automatic typing of dbpedia entities. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 65–81, Berlin, Heidelberg, 2012. Springer-Verlag.
7. A. Gangemi, V. Presutti, and D. R. Recupero. Frame-based detection of opinion holders and topics: A model and a tool. *IEEE Comp. Int. Mag.*, 9(1):20–30, 2014.
8. P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann. RelFinder: Revealing relationships in RDF knowledge bases. In *Proceedings of the 3rd International Conference on Semantic and Media Technologies (SAMT)*, volume 5887 of *Lecture Notes in Computer Science*, pages 182–187. Springer, 2009.
9. A. D. Iorio, A. G. Nuzzolese, and S. Peroni. Towards the automatic identification of the nature of citations. In A. G. Castro, C. L. 0002, P. W. Lord, and R. Stevens, editors, *SePublica*, volume 994 of *CEUR Workshop Proceedings*, pages 63–74. CEUR-WS.org, 2013.
10. A. Musetti, A. G. Nuzzolese, F. Draicchio, V. Presutti, E. Blomqvist, A. Gangemi, and P. Ciancarini. Aemoo: Exploratory search based on knowledge patterns over the semantic web. *Semantic Web Challenge*, 2012.
11. S. Peroni, A. Gangemi, and F. Vitali. Dealing with markup semantics. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 111–118, New York, NY, USA, 2011. ACM.
12. S. Peroni and D. Shotton. Fabio and cito: Ontologies for describing bibliographic resources and citations. *J. Web Sem.*, 17:33–43, 2012.
13. S. Peroni and D. Shotton. Fabio and cito: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:33–43, 2012.
14. V. Presutti, S. Consoli, A. G. Nuzzolese, D. Reforgiato Recupero, A. Gangemi, I. Bannour, and H. Zargayouna. Uncovering the semantics of wikipedia wikilinks. 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2014), 2014.
15. V. Presutti, F. Draicchio, and A. Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In *Knowledge Engineering and Knowledge Management*, volume 7603 of *Lecture Notes in Computer Science*, pages 114–129. Springer Berlin Heidelberg, 2012.
16. D. Reforgiato Recupero, V. Presutti, S. Consoli, A. Gangemi, and A. G. Nuzzolese. Sentilo: Frame-based sentiment analysis. *Cognitive Computation*, 2014.
17. M. van Assem, A. Gangemi, and G. Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006.