Extracting knowledge from text using SHELDON, a Semantic Holistic framEwork for LinkeD ONtology data

Diego Reforgiato Recupero, Andrea G. Nuzzolese, Sergio Consoli, Valentina Presutti, Silvio Peroni, Misael Mongiovì
National Research Council
ISTC/STLab
Italy
diego.reforgiato@istc.cnr.it, andrea.nuzzolese@istc.cnr.it, sergio.consoli@istc.cnr.it, valentina.presutti@istc.cnr.it, silvio.peroni@unibo.it, misael.mongiovi@istc.cnr.it

ABSTRACT

SHELDON[1] is the first true hybridization of NLP machine reading and the Semantic Web. It extracts RDF data from text using a machine reader: the extracted RDF graphs are compliant to Semantic Web and Linked Data. It goes further and applies Semantic Web practices and technologies to extend the current human-readable web. The input is represented by a sentence in any language. SHELDON includes different capabilities in order to extend machine reading to Semantic Web data: 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, nice visualization tools which make use of the JavaScript infoVis Toolkit and RelFinder. A demo of SHELDON can be seen and used at `http://wit.istc.cnr.it/stlab-tools/sheldon`.

Keywords

Semantic Web; Machine Reading; Linked Data

1. INTRODUCTION

In order to extract knowledge from text, the Machine Reading paradigm adopts Natural Language Processing (NLP) algorithms and methods. 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 [5]. SHELDON performs a hybrid (part of the components are trained, part are rule-based), self-supervised variety of machine reading tasks that generates RDF graph representations out of the knowledge extracted from text by dedicated NLP tools. The produced graphs represent an extension and improvement of NLP output, and can be customized to target specific tasks. Several software modules successfully evaluated in the recent past [14, 6, 4, 7, 17, 9, 13, 1, 8, 10, 16, 5] constitute the core components where SHELDON builds on top.

The machine reading capability of SHELDON is based on FRED [14, 4], a powerful component that automatically creates RDF/OWL graphs containing linked data from text. FRED integrates, transforms, improves, and abstracts the output of several NLP tools. Boxer [2] is a deep semantic parser that is called by FRED. Boxer includes a statistical parser (C&C) and generates 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). The latter generates formal semantic representation of text through an event (neo-Davidsonian) semantics. The basic NLP tasks performed by FRED by means of Boxer include: event detection (DOLCE+DnS[2] [3] is used by FRED), semantic role labeling with VerbNet[3] and FrameNet roles, first-order logic representation of predicate-argument structures, logical operator scoping (called boxing), modality detection, tense representation, entity recognition using TAGME[4], word sense disambiguation (the next version is going to use BabelNet[5]), DBpedia for expanding tacit knowledge extracted from text, etc. Everything is integrated and semantically enriched in order to provide a Semantic Web-oriented reading of texts. FRED is also accessible by means of a Python API, namely *fredlib*[6]. It exposes features for retrieving FRED graphs from user-specified sentences, and managing them.

Uncovering the semantic meaning of hyperlinks has a huge impact on the knowledge that can be extracted from Web data and that can therefore be published in machine readable form. The correspondence with the natural language source can be maintained and this would further benefit the acquired knowledge. SHELDON integrates LEGALO [13],

[1]This project receives funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 643808.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
`http://dx.doi.org/10.1145/2740908.2742842`.

[2]DOLCE+DnS Ultralite Ontology. `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`
[3]T. V. project. `http://verbs.colorado.edu/mpalmer/projects/verbnet.html`
[4]`http://tagme.di.unipi.it/`
[5]`http://babelnet.org/`
[6]`http://wit.istc.cnr.it/stlab-tools/fred/fredlib`

a novel method for revealing the semantics of links by identifying new semantic relations for them. Using a set of graph pattern-based heuristics, LEGALO extracts from FRED graphs Semantic Web binary relations that capture well the semantics of the underlying links.

SHELDON is able to give a boost to the sentiment analysis practices [15]. One of its components is built on top of SENTILO [17, 7], an independent sentic computing approach which uses both natural language processing techniques and Semantic Web technologies. For a certain sentence expressing an opinion, SENTILO is able to identify its holder, to extract the topics and subtopics that the holder targets, to link them to related events or situations present in the text and to evaluate the sentiment expressed on each topic/subtopic. SENTILO uses a new 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 analyses the Wikipedia page abstracts containing the natural language definition of entities and looks for the most appropriate type for them. TÌPALO relies on FRED for parsing and representing the logical form of a given sentence and induces a taxonomy by reusing WordNet types, WordNet supersenses, and DUL types.

Within academic communities, bibliographic citations earned a huge importance for linking scientific papers to related works, experiments, surveys, etc. SHELDON uses one more method to pursuit the task above. More in detail, SHELDON embeds CITALO [9], a tool that exploits Semantic Web technologies and NLP techniques to automatically infer the purpose of citations. The input is represented by CiTO [12], an ontology which describes the nature of citations in scientific research papers and a paragraph containing a reference to a bibliographic entity. 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, SHELDON provides a data exploration component, built on top of the Semantic Scout [1], which uses the JavaScript InfoVis Toolkit[8]. Finally, it is possible to get a much wider knowledge of the relations between detected DBpedia entities using a SHELDON component that is built upon the expansion algorithm described for RelFinder [8] and that shows further relations between the detected DBpedia entities.

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).

[7]Graphviz - Graph Visualization Software, `http://www.graphviz.org/`
[8]`http://philogb.github.io/jit/`

2. SHELDON AT WORK

Figure 1 shows the main interface of SHELDON where it is possible to insert some text in any language and select the semantic feature for processing the text. The output produced by SHELDON is in English no matter what the source language is. As the semantic core of SHELDON processes text using English, Bing Translation APIs[9] are used to translate input text (given in 47 possible languages) in English.

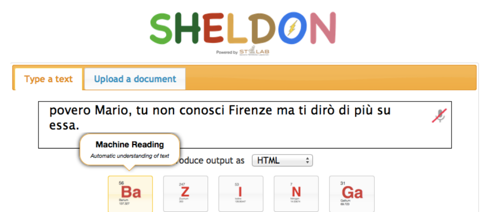

Figure 1: SHELDON front page

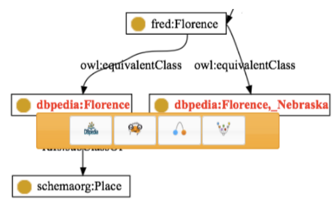

Figure 2: SHELDON's navigation toolbar for identified DBpedia entities

The input text might be given in any of 47 different languages. If it is not English, there is an automatic language detection module[10] that will identify the source language.

For example, the Italian sentence listed in Figure 1 would be correctly translated from :
"povero Mario, tu non conosci Firenze ma ti dirò di più su essa."
to:
"poor Mario, you don't know Florence but I'll tell you more about it."
SHELDON will work with the English representation of it.

Besides, using the Google Chrome browser it is possible to use the Google Web Speech API recognition feature[11] which would translate the spoken text in natural language text.

SHELDON addresses the following four main tasks:

- If the user chooses to perform machine reading (by pressing the button "Ba" of the toolbar - cf. Figure 1), an RDF output of the related text is shown. Several information such as detected DBpedia entities, events and situations mapped within DOLCE, WordNet and VerbNet mapping, pronoun resolution, and so on are displayed on the graph. Figure 4 shows a part of the graph produced by SHELDON for the example sentence and related to the machine reading component.

[9]`http://www.microsoft.com/web/post/using-the-free-bing-translation-apis`
[10]It uses the Language Identification Engine of Apache Stanbol.
[11]`https://www.google.com/intl/en/chrome/demos/speech.html`

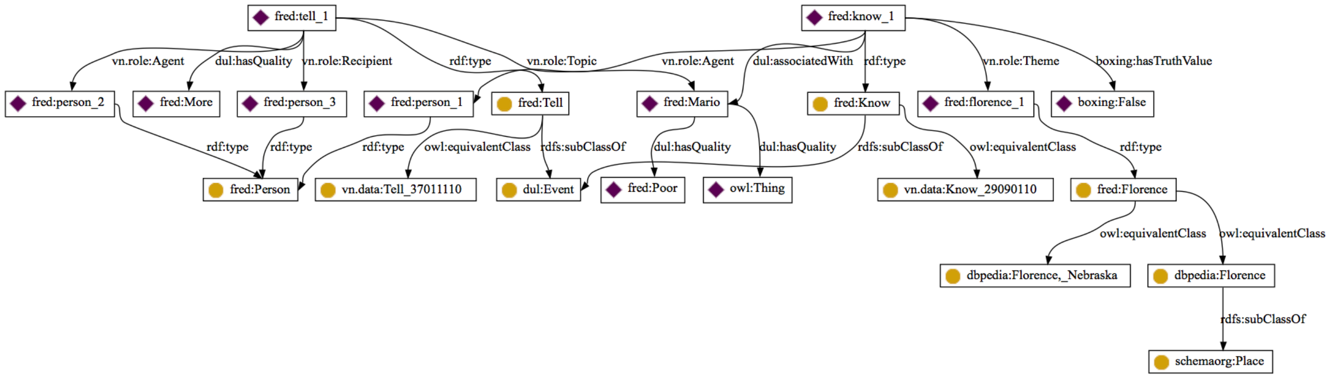

Figure 4: Machine reader output for the Italian text: *"povero Mario, tu non conosci Firenze ma ti dirò di più su essa."*

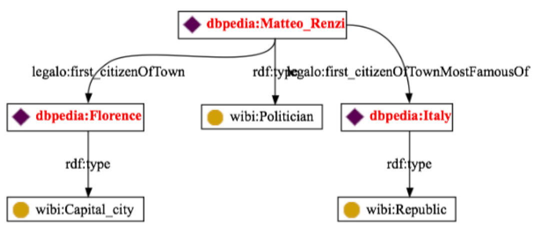

Figure 5: Semantic link identification for the French sentence *"Matteo Renzi a été le premier citoyen de l'une des villes plus célèbres de l'Italie, Florence."*

References

[1] Baldassarre, C., Daga, E., Gangemi, A., Gliozzo, A. M., Salvati, A., and Troiani, G. Semantic scout: Making sense of organizational knowledge. In *EKAW* (2010), P. Cimiano and H. S. Pinto, Eds., vol. 6317 of *Lecture Notes in Computer Science*, Springer, pp. 272–286.

[2] Bos, J. Wide-coverage semantic analysis with boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing* (Stroudsburg, PA, USA, 2008), STEP '08, Association for Computational Linguistics, pp. 277–286.

[3] Gangemi, A. What's in a schema? *Cambridge University Press, Cambridge, UK* (2010), 144–182.

[4] Gangemi, A. A comparison of knowledge extraction tools for the semantic web. In *ESWC* (2013), P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, Eds., vol. 7882 of *Lecture Notes in Computer Science*, Springer, pp. 351–366.

[5] Gangemi, A., Draicchio, F., Presutti, V., Nuzzolese, A. G., and Recupero, D. R. A machine reader for the semantic web. In *International Semantic Web Conference (Posters & Demos)* (2013), E. Blomqvist and T. Groza, Eds., vol. 1035 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 149–152.

[6] Gangemi, A., Nuzzolese, A. G., Presutti, V., Draicchio, F., Musetti, A., and Ciancarini, P. Automatic typing of dbpedia entities. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I* (Berlin, Heidelberg, 2012), ISWC'12, Springer-Verlag, pp. 65–81.

[7] Gangemi, A., Presutti, V., and Recupero, D. R. Frame-based detection of opinion holders and topics: A model and a tool. *IEEE Comp. Int. Mag. 9*, 1 (2014), 20–30.

[8] Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., and Stegemann, T. RelFinder: Revealing relationships in RDF knowledge bases. In *Proceedings of the 3rd International Conference on Semantic and Media Technologies (SAMT)* (2009), vol. 5887 of *Lecture Notes in Computer Science*, Springer, pp. 182–187.

[9] Iorio, A. D., Nuzzolese, A. G., and Peroni, S. Towards the automatic identification of the nature of citations. In *SePublica* (2013), A. G. Castro, C. L. 0002, P. W. Lord, and R. Stevens, Eds., vol. 994 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 63–74.

[10] Musetti, A., Nuzzolese, A. G., Draicchio, F., Presutti, V., Blomqvist, E., Gangemi, A., and Ciancarini, P. Aemoo: Exploratory search based on knowledge patterns over the semantic web. *Semantic Web Challenge* (2012).

[11] Peroni, S., Gangemi, A., and Vitali, F. Dealing with markup semantics. In *Proceedings of the 7th International Conference on Semantic Systems* (New York, NY, USA, 2011), I-Semantics '11, ACM, pp. 111–118.

[12] Peroni, S., and Shotton, D. Fabio and cito: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web 17* (2012), 33–43.

[13] Presutti, V., Consoli, S., Nuzzolese, A. G., Reforgiato Recupero, D., Gangemi, A., Bannour, I., and Zargayouna, H. Uncovering the semantics of wikipedia wikilinks. 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2014).

[14] Presutti, V., Draicchio, F., and Gangemi, A. Knowledge extraction based on discourse representation theory and linguistic frames. In *Knowledge Engineering and Knowledge Management*, vol. 7603 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 114–129.

[15] Presutti, V., Stankovic, M., Cambria, E., Cantador, I., Iorio, A. D., Noia, T. D., Lange, C., Recupero, D. R., and Tordai, A., Eds. *Semantic Web Evaluation Challenge - SemWebEval 2014 at ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers* (2014), vol. 475 of *Communications in Computer and Information Science*, Springer.

[16] Recupero, D. R., Consoli, S., Gangemi, A., Nuzzolese, A. G., and Spampinato, D. A semantic web based core engine to efficiently perform sentiment analysis. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers* (2014), pp. 245–248.

[17] Reforgiato Recupero, D., Presutti, V., Consoli, S., Gangemi, A., and Nuzzolese, A. G. Sentilo: Frame-based sentiment analysis. *Cognitive Computation, in press* (2014).