Bringing The Output of Open Information Extraction to The RDF/XML Format: A Case Study

Alisa Zhila[1] *, Elena Yagunova[2], and Olga Makarova[2]

[1] Instituto Politécnico Nacional,
[2] Saint Petersburg State University, eng.spbu.ru
{alisa.zhila,iagounova.elena,makarovaolgae}@gmail.com

Abstract. Open Information Extraction (OIE) is one of the most used strategies for Information Discovery and Knowledge Acquisition. However, the format of the output of OIE systems (although traditionally presented in the form of semi-structured tuples) has not been standardized and is defined solely by the design of a particular system. To successfully incorporate the OIE output into a target application such as Semantic Indexing or Ontology Population, its conversion to a standardized format is needed. In this paper we introduce a procedure for output conversion for a particular OIE system into the RDF/XML format. We address such issues as normalization of extracted components and conversion of multicomponent tuples into RDF triples. As a result, our procedure converts an extraction returned by the OIE system into an RDF graph. The resulting RDF/XML representation passes validation by the official W3C RDF/XML validator.

Keywords: open information extraction, information discovery, RDF/XML, standardization

1 Introduction

Open Information Extraction (OIE) from text is a widely used paradigm of IE and is a method for Information Discovery and Knowledge Acquisition. It is a necessary step for such target applications as Question Answering, Semantic Indexing, Semantic Search, Knowledge Base Population. OIE tradition suggests returning the extractions in a form of semi-structured tuples as in the example:

⟨*The BBC's correspondent Steve Rosenberg*⟩ ⟨*said*⟩ ⟨*the van*⟩ ⟨*belonged to*⟩ ⟨*Denis Pushilin*⟩

In fact, the tuple structure and number of components is defined arbitrarily by the designer of a particular OIE system. Yet any further use of extractions in a target application inevitably requires their presentation in a particular format. Although some tools work with custom formats, the common practice is to use one of the existing standardized data presentation formats. RDF/XML is a format of information presentation standardized by W3C (The World Wide Web Consortium) in [**?**] and widely used in Semantic Web and for Semantic Indexing in databases [**?**] among others.

In this work, we introduce a procedure for presentation of extractions returned by an OIE system described in [**?**] in the RDF/XML format. Our procedure solves such issues as normalization of the extracted components and conversion of multicomponent tuples into RDF statements.

2 Procedure Description and Validation

For the format conversion, we take the output of the OIE system described in [**?**] because this system includes preliminary post-processing of the extractions. It splits semantically complex components into simpler units and performs shallow semantic interpretation of relations between the components.

After the post-processing, each final component of the extraction can be considered as a node in a semantic graph. Yet to be considered an IRI (International Resource Identifier) as required by the RDF standard, it should be normalized. We suggest the following normalization procedure:

- eliminate all determiners;
- eliminate lexical prepositions (they were processed by the OIE system);
- delete spaces between proper nouns of named entities keeping all first letters capitalized: *Steve Rosenberg* ⇒ *SteveRosenberg*;

* The work was done during the internship in Oracle Mexico Development Center.

– delete spaces inside other multiword components converting the first word into lowercase and capitalizing all the following words: *Deputy Finance Minister* ⇒ *deputyFinanceMinister* or *is planning* ⇒ *isPlanning*.

Then we connect the normalized components in an RDF graph. By the W3C standard, an RDF statement is a triplet. However, a sentence may contain more complex relations. Therefore, they are split into triples. This is done by introducing custom "blank" nodes that do not correspond to any lexical expression but to a generic relation.

In the absence of a universal comprehensive RDF vocabulary, we suggest using a very high-level syntactic based vocabulary that includes the following: `verb`, `subject`, `object`, `objectEvent`, and corresponding prepositional relations. Using this vocabulary, the RDF graph that corresponds to the extractions from the sentence "*The BBC's correspondent Steve Rosenberg says the van belonged to Denis Pushilin.*" is depicted in Figure **??**.

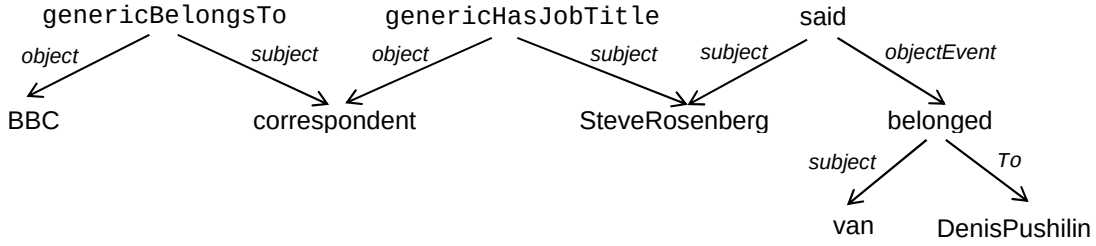

Fig. 1. An RDF graph corresponding to extractions from a sentence.

We have tested our procedure on a test set of 157 news articles from the Reuters News corpus [**?**] preliminary processed by the OIE system. The RDF/XML output of our conversion procedure passes validation test by the official W3C RDF/XML validator⋆. The RDF/XML representation can be constructed for extractions from any arbitrary text analyzed by the OIE method presented in [**?**].

3 Contributions

We introduced and tested a procedure for conversion of the OIE output into the RDF/XML format for a particular OIE system. This work has a practical value of bridging Information Discovery with such target applications as Semantic Indexing, Semantic Search, and others that use RDF representation. In particular we:

– stated the problem of the absence of a standard format for OIE output;
– proposed RDF/XML because it is a commonly used and standardized format;
– designed the procedure of OIE output conversion to RDF/XML for a particular OIE system;
– successfully tested the results with the official W3C RDF/XML validator.

In future, we will continue working on diffusion of RDF/XML as a standard format for OIE output; elaborate conversion procedures for more complex extractions; extend it for various OIE methods.

Acknowledgments. The authors acknowledge Saint Petersburg State University for research grant 30.38.305.2014. We also thank Dr. Souripriya Das for valuable insights and technical advice.

References

1. Gandon, F., Schreiber, G.: RDF 1.1 XML Syntax, `http://www.w3.org/TR/rdf-syntax-grammar`. W3C Working Group. Date of access (2014)
2. Lewis, D., Yang, Y., Rose, T., Li, F.: RCV1: A New Benchmark Collection for Text Categorization Research. J. Mach. Learn. Res. 5, pp. 361–397. JMLR.org (2004)
3. Oracle: Oracle Spatial and Graph: Advanced Data Management, `http://www.oracle.com/technetwork/database/options/spatialandgraph/overview/spatial-and-graph-wp-12c-1896143.pdf` (2014)
4. Zhila, A., Gelbukh, A., Gomez-Adorno, H.: Fast Named Entity Driven Open Information Extraction with Shallow Semantic Interpretation. Information Sciences. *Submitted* (2015)

⋆ `http://www.w3.org/RDF/Validator/`