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

Alisa Zhila1, Elena Yagunova2, Olga Makarova2
1Independent researcher, 2St. Petersburg State University
{alisa.zhila, iagounova.elena, makarova.olga.e}@gmail.com

Open Information Extraction
- Introduced by Michele Banko et al. in 2007
- Strategy for Information Discovery in arbitrary texts
- Arbitrary relations are numerous. It is not possible to make an exhaustive list of all relations and their arguments
- Traditional Information Extraction (IE) methods focus on particular relations
- OIE methods extract arbitrary relations based on specific NLP patterns without requiring a pre-specified vocabulary or large manually tagged training corpora
- Relations are extracted in the form of tuples

Example:
"Correspondent Steve Roizner works for the BBC"
Arg1: works
Rel: hasJobTitle
Arg2: the BBC

Potential Applications
- Knowledge Base
- Semantic Search
- Ontologies
- Others

Problems:
- No standard output format for OIE systems
- No general approaches on how to represent complex language phenomena, e.g., reported speech

Case-study OIE system
We chose an OIE system with the most extended semantic interpretation of extracted relations
Named-Entity Driven OIE System (Zhila et al. 2015)

Conversion to RDF
I. RDF Vocabulary: Properties
We use a syntax-based vocabulary: verb, subject, object, objectEvent, prepositions.

II. Correspondences with RDF model
1) Extracted tuple components ↔ RDF literals
   e.g., <belonged> ↔ "belonged", <the van> ↔ "van"
2) Non-verb based relations are represented as semantically meaningful nodes, e.g., jobTitle

Example:
Correspondent Steve Roizner said the van belonged to Denis Pushilin.

1. Part-of-speech tagging
   Correspondent*(NN) Steve*(NNP) Roizner*(NNP) said*(VBD) the*(DT) van*(NN) belonged*(VBD) to*(TO) Denis*(NNP) Pushilin*(NNP)

2. Named entity recognition
   Correspondent [Steve Roizner] Person said the van belonged to [Denis Pushilin] Person

3. Syntactic chunking

4. Information extraction
   <Correspondent Steve Roizner> said the van belonged to Denis Pushilin

5. Post-processing rules
   <Steve Roizner> hasJobTitle correspondent

Steps 1-3 are performed with GATE NLP Toolkit
Steps 4, 5 are the proper OIE system

Output represented as a semantic graph

III. Uniqueness
To make the event, eventSpeech and jobTitle nodes unique, we enumerate them throughout the document

IV. RDF/XML representation
We chose this format for convenience in a particular task.
All RDF formats are equivalent.

V. Validation
We successfully validated the output with the official W3C RDF/XML validator at http://www.w3.org/RDF/Validator/

Conclusions
- Conversion procedure for OIE output to RDF model
- Universal syntax-based property vocabulary
- Arbitrary verb-based relations represented in RDF through abstract event nodes
- Reported speech representation via eventSpeech - event
- Uniqueness throughout the document

We acknowledge Saint Petersburg State University for research grant 30.38.305.2014.