OPEN INFORMATION EXTRACTOR FOR SEVA

Jitin Krishnan
Mentor: Patrick Coronado

Other Project Members:
Trevor Reed (NASA JPL)
Ben Brumback (George Mason University)
Youngyo Na (Rowan University)
SEVA: A Systems Engineer’s Virtual Assistant

Systems Engineer

- Question-Answering
- Explainable System
- Dynamic Information
- Time & Scheduling
- Hypothetical Scenarios
- Capture Experience

Personal Assistant

Information Assimilation

Documents
(Eg: SE Handbook)
**SEVA: ARCHITECTURAL COMPONENTS**

**Motivation**
- Information Assimilation
- Explainability
- Personal System
- Domain-specific Common Sense

**INPUT**
- QUESTION
- FACTS/STATEMENTS
- SYSTEM COMMANDS

**OUTPUT**
- ANSWER
- QUESTION*
- SYSTEM RESPONSES

**NATURAL LANGUAGE PROCESSING UNIT**
- TARGETTED OPEN INFORMATION EXTRACTION
- DICTIONARY & VERB NORMALIZATION
- RULE EXTRACTION
- QUERY LANGUAGE

**DYNAMIC KNOWLEDGE BASE** (WITH TEMPORAL RELATIONS)
- BASE KB + RULES
- CUSTOM KB + RULES

**INFERERENCE ENGINE**
- ONTOLOGY LANGUAGE REASONER
- RULE REASONER
- TEMPORAL REASONER

**HYPOTHETICAL MODE**

**CASE-BASED EXPERIENCE**

*in conversational mode (rule/experience building)
SEVA: SYSTEM CONCEPT

- Interactive Q&A
- Natural Language Processing
- Inference Engine
- Dynamic Ontology
- Monitor
- Logicker
- Endowed Models
- Inter-Module Communication Protocol
**SEVA:** Knowledge Representation Example

**Sentence:** Instruments [have] thermal zones

Subject: **Instruments**  
Verb/Predicate: **have**  
Object: **thermal zones**

**First Order Logic (FOL) Formalism:**

\[ \forall x \ \exists y \ \text{Instrument}(x) \land \text{Thermal Zone}(y) \Rightarrow \text{have}(x,y) \]

**Concepts:** Instrument, Thermal Zone (like classes in OOP)  
**Instances:** x, y  
**Relationship:** have
**SEVA: Knowledge Representation Example**

**ABox**

*Instrument*(STI)
STI is an instrument

*part of*(MassSpectrometerAP8717, STI)
MassSpectrometerAP8717 is a part of STI

**TBox**

*Conduit ≡ Pipe*
Conduit is same as Pipe

*MassSpectrometer ⊑ Spectrometer*
Subclass relationship

**RBox**

*partOf ◦ partOf ⊑ partOf*
Transitive property of the role

*partOf ⊑ hasComponent ~*
Inverse property of two roles
**Project Specific Documents**

STI is an instrument. STI has a length of 200 cm. Mass Spectrometer MS81Z is a part of STI.

**SE Handbook, Common-Sense Knowledge**

Conduit is same as Pipe. Mass Spectrometer is a type of Spectrometer. Instruments have thermal zones.

**Verb-oriented Common-Sense Knowledge**

Verb Vocabulary, Word Net, DBPedia, Word2Vec contextual synonyms

---

**OIE**

[STI; is-a; instrument] [STI; has; length] [length; has-value; 200 cm] [Mass Spectrometer MS81Z; part of; STI]

[Conduit; same as; Pipe] [Mass Spectrometer; type of; Spectrometer] [Instruments; have; thermal zones]

[partOf: transitive] [partOf; inverse-of; hasComponent]
Open Information Extraction

• Identifies wide range of domain-independent relations
• Traditional Information Extraction: uses predetermined templates

Examples: Stanford Open IE, Open IE by AI2, ClausIE

• For the experiment the domain language complexity is reduced – we work with only simple English sentences
**NLP Basics**

*Sentence:*
STI, an instrument, weighs 56 kg

**Dependency Parse Tree**

```
weighs
   |
  STI
  |
instrument
   |
    an
    |
      kg
```

**Part-of-Speech:**

- **NN** = Noun
- **VBZ** = Verb form

```
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>DT</td>
<td>NN</td>
<td>VBZ</td>
<td>CD</td>
</tr>
</tbody>
</table>
```

1. STI, an instrument, weighs 56 kg.

**Basic Dependencies:**

```
NN   punct
punct appos
nsubj
```

**Pattern Matching on the dependency tree**

**Extracting relations of type:**
- *is-a*, transitive-verb, has-property, has-value

**A set of universal dependencies:**
- nsubj, dobj, case, nmod, compound, amod

**Chunking example:**

"NP:{(<CD.*>|<JJ.*>)<NN.*>+}"

**SEVA – Targeted Open IE (TOIE)**

**NLTK, Stanford CoreNLP, POS Tagger**
<table>
<thead>
<tr>
<th>Input Sentence</th>
<th>Stanford Open IE</th>
<th>ClausIE</th>
<th>A12 Open IE</th>
<th>SEVA-TOIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>STI, an instrument, has a 2500 pixel CCD detector</td>
<td>(“STI” “has” “2500 pixel CCD detector”)</td>
<td>(“STI” “is” “an instrument”)</td>
<td>(STI; has; a 2500 pixel CCD detector)</td>
<td>(STI; has; CCD detector)</td>
</tr>
<tr>
<td></td>
<td>*incomplete/missing information: “STI is an instrument”</td>
<td>(“STI” “has” “2500 pixel CCD detector”)</td>
<td>(STI; has; CCD detector)</td>
<td>(CCD detector; has-property; 2500 pixel)</td>
</tr>
<tr>
<td>STI is an instrument with a TRL value of 5</td>
<td>(“STI” “is” “instrument”)</td>
<td>(“STI” “is” “an instrument with a TRL value of 5”)</td>
<td>(STI; is; an instrument with a TRL value of 5)</td>
<td>(STI; is; instrument)</td>
</tr>
<tr>
<td></td>
<td>(“STI” “is” “instrument with TRL value of 5”)</td>
<td>(“STI” “is” “an instrument with a TRL value of 5”)</td>
<td>(STI; is an instrument with a TRL value of 5)</td>
<td>(instrument; has-property; TRL value)</td>
</tr>
<tr>
<td></td>
<td>(“STI” “is instrument with TRL value”)</td>
<td>(“STI” “is” “an instrument with a TRL value of 5”)</td>
<td>(STI; is an instrument with a TRL value of 5)</td>
<td>(TRL value; has-value; 5)</td>
</tr>
<tr>
<td>STI is scheduled for acoustic testing on July 3, 2015 from 2:00PM to 6:00PM.</td>
<td>produced 13 triples</td>
<td>(“STI” “is scheduled for acoustic testing on July 3 2015 from 2:00 PM”)</td>
<td>(STI; is scheduled; for acoustic testing)</td>
<td>(STI; is; scheduled)</td>
</tr>
<tr>
<td></td>
<td>(“STI” “is” “scheduled”)</td>
<td>(“STI” “is scheduled for acoustic testing on July 3 2015 from 2:00 PM”)</td>
<td>incomplete/missing information: “on July 3, 2015”</td>
<td>(scheduled; for; testing)</td>
</tr>
<tr>
<td></td>
<td>(“STI” “is” “scheduled”)</td>
<td>(“STI” “is scheduled for acoustic testing on July 3 2015 from 2:00 PM”)</td>
<td>from: 2:00 PM</td>
<td>(testing; has-property; acoustic)</td>
</tr>
<tr>
<td></td>
<td>X = various combinations of remaining sentence</td>
<td>(“STI” “is scheduled for acoustic testing on July 3 2015 from 2:00 PM”)</td>
<td>(scheduled; from; 2:00 PM)</td>
<td>(testing; has-value; 3 2015 July)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(“STI” “is scheduled for acoustic testing on July 3 2015 from 2:00 PM”)</td>
<td>(scheduled; to; 6:00 PM)</td>
<td>(scheduled; to; 6:00 PM)</td>
</tr>
</tbody>
</table>

Table 1: Output from Open Information Extractors. *Red* colored extractions are incomplete, incorrect, or noisy; *Blue* colored extractions need to be further granularized for ontology population.
THANK YOU!